Nervenarzt 2007 · 78:1028–1036
DOI 10.1007/s00115-007-2308-y
Online publiziert: 10. August 2007

W. Maier[1] · H.-J. Möller[2]
[1] Klinik für Psychiatrie und Psychotherapie der Rheinischen Friedrich-Wilhelms-Universität, Bonn
[2] Klinik für Psychiatrie und Psychotherapie der Ludwig-Maximilians-Universität, München

Metaanalysen

Methode zur Evidenzmaximierung von Therapiestudien?

Es ist weithin akzeptiert, dass Therapieempfehlungen und -entscheidungen evidenzbasiert sein sollten. Evidenz stützt sich auf empirische Studien und beansprucht Objektivität, indem alle informativen Untersuchungen gleichermaßen Berücksichtigung finden [13]. Grundlage der evidenzbasierten Medizin (EBM) ist die Kombination der Evidenzen aus den verschiedenen Einzelstudien und deren Zusammenführung zu einer umfassenden Schlussfolgerung. Systematische Übersichtsarbeiten [3] stellen Evidenzen und Schlussfolgerungen für spezifische Fragestellungen dar.

Die Aussagen verschiedener Studien sind aber oft uneinheitlich. Daher ist die Kumulation von Einzelevidenzen gemäß einem nachvollziehbaren Rationale durch Anwendung von Evidenzkriterien bzw. -stufen nötig; die Regeln dieses Verfahrens sind aber nicht natürlicherweise gegeben, sondern sie sind erst durch Konventionen festzulegen. Zu einer allgemeinverbindlichen Festlegung von Evidenzkriterien ist es bislang nicht gekommen; vielmehr werden unterschiedliche Bewertungskriterien vorgeschlagen [18].

Prinzipiell werden zwei Grundformen systematischer Evidenzgewinnung und -darstellung unterschieden:

1. Vergleichende und kumulative Wertung von Einzeluntersuchungen in *narrativen Übersichtsarbeiten*, aber ohne eine spezifische statistische Technik der Kombination von Einzelevidenzen; die methodisch am weitesten entwickelte Form ist dabei das *systematische Review*, in dem die Gesamtheit verfügbarer Studien mit jeweils methodenkritischer Wertung einzelner Studien (im Hinblick auf interne und externe Validität, [14]) und mit Wichtung der Besonderheiten einzelner Studien (z. B. im Hinblick auf Setting, Therapieziel) vergleichend dargestellt wird; hieraus werden qualitative Schlussfolgerungen abgeleitet (z. B. zur Wirkung/Wirksamkeit). In der Regel werden solche Schlussfolgerungen mit Caveats versehen (z. B. Evidenzlage ist noch unzureichend o. ä.).
2. Kumulative Wertung von Einzeluntersuchungen in Metaanalysen, wobei statistische Methoden für die Kombination von Einzelevidenzen genutzt werden; Metaanalysen gehen eine systematische Suche nach informativen Studien und deren methodenkritische Wertung voraus; die methodische Qualität von Studien kann dabei durch standardisierte Bewertungsinstrumente (Skalen) beurteilt werden. In die abschließende kumulative Wertung gehen nur qualitativ hinreichende Studien ein; dieses Verfahren kommt zu abschließenden quantitativen Wertungen, die auch auf Signifikanz geprüft werden können.

Die erste Strategie kann Besonderheiten spezifischer Einzelstudien und auch konzeptuelle Aspekte berücksichtigen; die mögliche Subjektivität des Autors v. a. bei der Globalbewertung wird aber auch als Einschränkung angeführt [3]. Metaanalysen sind dagegen auf spezifische Prüfhypothesen fokussiert. Alle eingeschlossenen Studien werden prinzipiell gleich behandelt, diese Annahme kann in sog. Sensitivitätsanalysen geprüft werden. Bei der Ableitung des Globaleffekts wird von der quantitativen Wichtung von Studien nach der methodischen Güte abgeraten [14].

Es besteht eine wachsende Tendenz, Metaanalysen die Priorität einzuräumen; so bildet „Evidenz aufgrund von Metaanalysen randomisierter kontrollierter Studien“ die höchste Evidenzstufe der einflussreichen „Agency of Health Care Policy and Research“ [5]. Wegen der formalisierten Durchführung und dem (scheinbar) eindeutigen statistischen Resultat stellen Metaanalysen zunehmend die Basis für Behandlungsempfehlungen in Leitlinien dar. Hierbei müssen jedoch die Grenzen der Aussagekraft von Metaanalysen beachtet werden. Trotz hoher Akzeptanz droht folglich eine – in der Verkennung der Begrenztheit der Aussagekraft – missbräuchliche Anwendung bzw. Verwendung von Metaanalysen. Wir führen im Weiteren fünf kritische Argumente für die Beschränktheit der Aussagekraft von Metaanalysen an.

Hier steht eine Anzeige.

Springer

Methodisches Vorgehen von Metaanalysen

Metaanalysen kombinieren die quantitativen Ergebnisse verschiedener Untersuchungen zur selben Fragestellung und summieren diese Ergebnisse in einer globalen Größe (z. B. Wirksamkeit von A im Vergleich zu Plazebo in einem definierten Zeitfenster). Einzelstudien zum Vergleich zweier oder mehrerer Therapiebedingungen prüfen a priori spezifizierte Hypothesen und sollten hierfür ausreichend Power (in Form von Stichprobenumfängen) haben. Die statistische Analyse erfolgt anhand von p-Werten, die Hypothesen verwerfen oder nicht verwerfen können; p-Werte sind aber stark von den Bedingungen der Einzelstudien (v. a. von der Hypothesenformulierung und den Stichprobenumfängen) abhängig und können daher nicht zwischen Einzelstudien vergleichen und damit auch nicht über die Studien kombiniert werden. Studienübergreifende, vergleichbare Ergebnisgrößen sind stattdessen die sog. Effektgrößen, z. B. die normierte Differenz zwischen zwei Vergleichsbedingungen; diese können unter Berücksichtigung der jeweiligen Power/Stichprobenumfänge in eine gemeinsame, alle eingeschlossenen Studien umfassende Effektgröße kombiniert werden. Für diese resultierende globale Effektgröße ist ein Konfidenzintervall ermittelbar, woraus sich auch Aussagen zur Signifikanz von Effekten ableiten. So kann auch bei inkonsistenter Ergebnislage eine umfassende globale quantitative Bewertung abgegeben werden, die alle relevanten Aussagen von Einzelstudien vollständig integriert.

Diese a priori plausible Strategie zur Zusammenfassung empirischer Evidenz erfordert die Spezifikation der zu entscheidenden Fragestellung. Die Fragestellung von Einzelstudien lässt sich retrospektiv aber nicht mehr normieren; die Gleichheit von Fragestellungen ist manchmal nicht eindeutig entscheidbar, so dass Studien einer Metaanalyse zugeordnet werden können, die eine ähnliche, aber nicht dieselbe Fragestellung behandeln. Ein Beispiel hierfür ist eine häufig zitierte Metaanalyse zur Rezidivprophylaxe mit Antipsychotika von Leucht et al. [15]; hier werden Studien zur Erhaltungs-

Zusammenfassung · Summary

Nervenarzt 2007 · 78:1028–1036 DOI 10.1007/s00115-007-2308-y

W. Maier · H.-J. Möller

Metaanalysen. Methode zur Evidenzmaximierung von Therapiestudien?

Zusammenfassung

Evidenzbasierte Medizin (EBM) entwickelt sich zur Richtschnur für klinische Therapieentscheidungen. Evidenz ist aber kein eindeutig in Kriterien explizierbarer Begriff. Es gibt viele verschiedene Evidenzkriterien. Metaanalysen werden dabei als ein vorrangiges Evidenzkriterium angesehen, da sie die möglicherweise unterschiedlichen Einzelevidenzen aus verschiedenen Studien zur selben Fragestellung zu einer globalen Bewertung kumulieren. Die Aussagekraft von Metaanalysen ist jedoch begrenzt. Hierzu werden fünf Argumente angeführt: (1) Die qualitativen Schlussfolgerungen aus der metaanalytischen Kombination einzelner Studien können von den methodischen Rahmenbedingungen der Studien abhängen; (2) die nötige Homogenität der Effektstärken der eingeschlossenen Studien ist oft nicht gegeben, was auch nur unzulänglich kontrolliert werden kann; (3) Metaanalysen betrachten die Variation von Ergebnissen der eingeschlossenen empirischen Untersuchungen als Störvarianz; die Variabilität von Ergebnissen über Studien kann aber informativ sein; (4) das Resultat von Metaanalysen ist abhängig von der Art, wie die eingeschlossenen empirischen Untersuchungen aufgefunden werden; (5) Schlussfolgerungen von Metaanalysen hängen qualitativ von der Auswahl der Teststatistik ab, die zur Kombination der Ergebnisse von Einzelstudien benutzt wird. Metaanalysen stellen vor allem eine Methode zur Hypothesengenerierung im Rahmen einer A-posteriori-Analyse von Therapieeffekten dar.

Schlüsselwörter

Metaanalysen · Evidenzmaximierung · Therapieentscheidung · Hypothesengenerierung · Teststatistik

Meta-analyses. A tool for maximizing therapy study evidence?

Summary

Medical decisions have to rely on evidence-based medicine. However, evidence is not a clearly defined term. Several evidence criteria have been proposed, yet statistical meta-analyses are considered to be core criterion of empirical studies in that they combine single evidence from different studies into an encompassing conclusion. We argue however that meta-analyses have major limitations in this context. Five arguments are presented: (1) the nature of a study design affects the results and thus comprehensive meta-analyses covering studies with different designs are not informative; (2) combining different studies into a meta-analysis is methodologically not informative; (3) meta-analyses consider the variance between studies not as informative but as random noise; (4) the strategy to identify informative studies is a decisive determinant of meta-analyses; and (5) the value of conclusions from meta-analyses depends on the choice of statistics chosen as combined results from the various single studies. Instead, meta-analyses are useful tools for generating hypotheses in a posteriori analysis.

Keywords

Forming a hypothesis · Maximizing evidence · Meta-analyses · Therapy decision · Test statistics

therapie mit solchen zur Rezidivprophylaxe kombiniert, obwohl es sich dabei lediglich um ähnliche, nicht aber um identische Fragestellungen handelt. Entsprechend erfordern Studien zur Erhaltungstherapie ein anderes Design als solche zur Rezidivprophylaxe.

Methodische Fortschritte

Metaanalysen sind mit einem schwer lösbaren Dilemma konfrontiert: Sie zielen einerseits auf Vollständigkeit, wobei alle zu einer Fragestellung verfügbaren Studien bzw. empirischen Resultate berücksichtigt werden sollen. Andererseits ist die methodische Qualität vorliegender Studien häufig heterogen, wobei einzelne Studien erhebliche methodische Insuffizienzen aufweisen können. Dieses Dilemma kann nur pragmatisch gelöst werden:

- Benennung von methodischen Minimalkriterien und
- Prüfung der zu einem Thema verfügbaren empirischen Untersuchungen auf Erfüllung dieser Minimalkriterien.

Die Metaanalyse wird dann auf die methodisch suffizienten empirischen Untersuchungen beschränkt.

Methodische Fortschritte erlauben es, die getroffene Aussagekraft von Metaanalysen zu erhöhen: Sensitivitätsanalysen können die Konsistenz der Aussagen charakterisieren oder Funnel-Plots können selektive Publikationen bzw. eine verfälschte Auswahl von Studien entdecken; die Ergebnishomogenität zwischen eingeschlossenen Studien kann durch „Heterogenitätstests" gesichert werden. Diese methodischen Verfeinerungen konnten die wahrgenommene Akzeptanz und Überzeugungskraft von Metaanalysen zusätzlich erhöhen.

Trotz dieser methodischen Fortschritte ist einschränkend auf eine Unstimmigkeit hinzuweisen:

Aussagekräftige Vergleichsstudien zur klinischen Wirksamkeit zielen heute auf eine qualitative Hypothese (A ist B überlegen) und sind entsprechend geplant (inkl. Power-Analysen zur Abschätzung des Stichprobenumfangs). Metaanalysen extrahieren quantitative Effektgrößen aus den Einzelstudien, die eigentlich auf einer qualitativen Fragestellung ausgerichtet sind. Metaanalysen testen also keine Hypothesen auf der Grundlage einer konsequenten Versuchsplanung; die resultierende Effektgröße sagt auch nichts über die Anzahl berücksichtigter Studien und Stichprobenumfänge aus. So besteht z. B. die Gefahr, dass klinisch nicht relevante Differenzen zwischen Vergleichsbedingungen aufgrund zu großer Stichprobenumfänge überinterpretiert werden; ebenso können auf schwacher empirischer Grundlage zufällig eindrucksvolle Effektgrößen resultieren, die fälschlicherweise Grundlage von Behandlungsempfehlungen werden können. Auch sind die kombinierten Einzelstudien nicht notwendigerweise nach denselben Prinzipien geplant und damit nicht automatisch austauschbare Wiederholungen von Prüfungen derselben Hypothese.

Vorausgesetzt wird eine unverfälschte Auswahl von Studien, was durch Funnel-Plots geprüft werden kann. Funnel-Plots haben aber nur eine begrenzte Trennschärfe, vor allem bei Einzelstudien mit geringerem Stichprobenumfang. Ebenso wird die Vergleichbarkeit von Studien vorausgesetzt, was durch Heterogenitätstestung geprüft werden kann. Die Trennschärfe von Heterogenitätstests ist jedoch so begrenzt, dass vorhandene Inkonsistenzen nicht immer mit hinlänglicher Sicherheit entdeckt werden können [6]. In beiden Fällen wird dann bei Nichtverwerfung der Hypothese auf das Fehlen eines Publikationsbias oder das Fehlen von Heterogenität geschlossen.

Kontroverser Status von Metaanalysen

Das „metaanalytische Design" [5] hat bereits abgeschlossene Studien zur selben Fragestellung zum Gegenstand; die Designs der eingeschlossenen Studien wurden bei ihrer Planung nicht aufeinander abgestimmt, ihr Design wurde insbesondere nicht im Hinblick auf spätere Metaanalysen entwickelt; die Einzelstudien wurden meist vielmehr aufgrund unterschiedlicher Kriterien geplant und weisen Unterschiede in Stichprobengenerierung und -struktur sowie in der Versuchsdurchführung auf. Über diese Differenzen können auch vergleichbare Einschlusskriterien und gleiche Erfolgsmaße nicht hinwegtäuschen. Mit anderen Worten, Metaanalysen arbeiten mit einem retrospektiven Design, im Gegensatz zu den prospektiv geplanten Einzelstudien. Aus diesem Grund ist die Methode „Metaanalyse" der herben Kritik führender Biometriker ausgesetzt (z. B. [8]).

Insbesondere gehen die Metaanalysen nicht von einer vorausgehenden Power-Analyse aus. Die Vereinigung aller Stichproben der Einzelstudien resultiert meist in einem so umfassenden Umfang, dass die Analysen „über-powered" sind, d.h. auch klinisch nicht relevante Wirksamkeitsdifferenzen werden mit hoher Wahrscheinlichkeit als signifikant nachgewiesen.

Metaanalysen können daher zur Irritation beitragen [19]:

- Damit können einerseits klinisch irrelevante Globaleffekte in Metaanalysen als statistisch signifikant erscheinen.
- Andererseits können methodisch insuffiziente Studien Verum-Placebo-Differenzen verdecken, so dass die Größe des wahren Effekts unterschätzt wird.

Diese Konstellation wurde z. B. aufgrund von zu geringen globalen Effekten als Argument gegen die relevante Wirksamkeit der Serotoninwiederaufnahmehemmer (SSRI) diskutiert [19]. Allein die kritische Bewertung jeder einzelnen Studie kann der unterschiedlichen methodischen Qualität Rechnung tragen (wie sie z. B. durch Zulassungsbehörden vollzogen wird).

Einige Expertengremien, die mit der kumulativen Bewertung von empirischer Evidenz zu einer Fragestellung befasst sind, orientieren sich nämlich an einem alternativen Rationale. Sie bewerten die Ergebnisse jeder einzelnen Studie vor dem Hintergrund der jeweiligen Qualität; eine globale Schlussfolgerung auf die Evidenzlage erfolgt dann – unter Berücksichtigung der Limitationen einzelner Studien – durch Gegenüberstellung von Studien mit positiven oder negativen Ergebnissen, ohne dass in einer Metaanalyse die quantitativen Einzelergebnisse in eine qualitative Globalgröße zusammengefasst werden. Eine globale Bewertung (im Sinne einer Behandlungsempfehlung) erfolgt nur bei eindeutiger Evidenzlage; hierfür werden qualitative Kriterien angegeben (s. z. B.

[4]). Nach diesem letztgenannten Rationale gehen z. B. alle nationalen und internationalen Zulassungsbehörden für Arzneimittel vor ebenso wie Leitlinienkommissionen nationaler und internationaler Fachgesellschaften (z. B. Weltverband für Biologische Psychiatrie; [4]).

Der Begriff „Evidenz" oder „empirische Evidenz" suggeriert Schlüssigkeit und Eindeutigkeit. Wie eben skizziert, sind aber verschiedene rationale Wege zur Generierung von empirischer Evidenz auf der Grundlage verschiedener empirischer Studien möglich. Bei inkonsistenter Ergebnislage der Einzelstudien können unterschiedliche Schlussfolgerungen resultieren. Welcher dieser beiden paradigmatischen Wege ist in Bezug auf die praktische Schlussfolgerung am überzeugendsten? Eine eindeutige Antwort auf diese Frage wird nicht möglich sein. Es können aber verschiedene Argumente dafür genannt werden, dass die Aussagekraft von Metaanalysen begrenzt ist.

Einschränkungen der Aussagekraft von Metaanalysen

Argument 1

Die qualitativen Schlussfolgerungen aus der metaanalytischen Kombination einzelner Studien können von den methodischen Rahmenbedingungen der Studien abhängen.

Metaanalysen kombinieren Studien, die methodisch unterschiedliche Voraussetzungen haben. Die Schlussfolgerungen aus Metaanalysen sind also damit möglicherweise von dem relativen Gewicht der überwiegend gewählten Untersuchungsmethoden abhängig. Jüni et al. [14] vergleichen metaanalytische Resultate in Abhängigkeit von mehreren methodischen Rahmenbedingungen (z. B. Adäquanz der Generierung und der Verblindung der Randomisierung, Handhabung von Studienabbrüchen). Inadäquanz in der Handhabung dieser Designkriterien führte jeweils zu Ergebnisverfälschungen. Für Wertung von Studien mit solchen Defiziten in Metaanalysen gibt es keine eindeutigen Lösungen. Jedenfalls sollten die Mängel jeder spezifischen Studie in qualitativer und quantitativer Hinsicht (wofür Checklisten zur Verfügung stehen) individuell bewertet werden; ihr verfälschender Einfluss kann in geeigneten Sensitivitätsanalysen zumindest deutlich gemacht werden [14]. Die Ergebnisse der einschlägige Untersuchungen kombinierenden Metaanalysen und die qualitativen Schlussfolgerungen variierten systematisch mit der methodischen Rahmenbedingung.

Argument 2

Metaanalysen gehen von einer weitgehenden Vergleichbarkeit der ausgewählten Studien zu einer Prüfhypothese aus. Die hierzu nötige Homogenität der Effektstärken der eingeschlossenen Studien ist oft nicht gegeben.

Metaanalysen gehen von der Annahme aus, dass ein (allen Einzeluntersuchungen) gemeinsamer Populationsparameter geschätzt wird; alle zu einer Prüfhypothese vorliegenden Studien werden als zufällig gezogene Realisierungen eines umfassenden Experiments in derselben Grundgesamtheit betrachtet. Diese Voraussetzung kann nicht selbstverständlich angenommen werden, denn jede Studie wird unter spezifischen Bedingungen geplant und durchgeführt. Studienergebnisse hängen möglicherweise nicht nur von der Prüfhypothese, sondern auch von den spezifischen Studienbedingungen ab.

Daher kann die genannte Annahme nicht einfach unterstellt werden; die Homogenität bzw. fehlende Heterogenität ist erst noch (vor Durchführung der Metaanalyse) zu prüfen. Allein durch das Auswahlkriterium methodischer Minimalbedingungen wird die nötige Ergebnishomogenität noch nicht gewährleistet. Die Prüfung der Homogenität der Studienergebnisse (d.h. der Effektstärken) kann durch geeignete Tests erfolgen. Diese haben meist eine geringe Teststärke und plädieren damit bevorzugt für Homogenität ([6], S. 637). Der verfälschende Einfluss von mangelnder Homogenität (Heterogenität) der Effektgrößen kann durch die Wahl eines geeigneten statistischen Modells („random-" statt „fixed-effects models") abgemildert, nicht aber beseitigt werden. Folgerichtig wird im Fall eines signifikanten „Heterogenitätstests" dafür plädiert, keine Metaanalyse durchzuführen oder diese auf geeignete, homogene Teilmengen von Untersuchungen zu beschränken ([6], S. 637).

Dieses Problem ist bei der Evaluation von Psychopharmaka virulent. So zeigen sog. Heterogenitätstests für den Wirksamkeitsvergleich von neueren zu klassischen Antipsychotika keine ausreichende Homogenität an [15]; diese Einschränkung der Aussagekraft findet jedoch in den Schlussfolgerungen keine entsprechende Berücksichtigung; sie wird gar nicht diskutiert!

Argument 3

Metaanalysen betrachten die Variation von Ergebnissen der eingeschlossenen empirischen Untersuchungen als Störvarianz. Die Variabilität von Ergebnissen über Studien kann aber informativ sein.

Metaanalysen betrachten die Ergebnisse der eingeschlossenen Studien als Wiederholungen der Prüfung derselben Prüfhypothese. Qualitativ unterschiedliche Ergebnisse empirischer Studien können möglicherweise auf unterschiedliche klinische Bedingungen zurückgeführt werden. Das Wissen um diese Varianzquellen kann erhebliche praktische Konsequenzen haben und sollte nicht vernachlässigt werden. Ein schlüssiges Beispiel stellt die Metaanalyse der antidepressiven Effekte von SSRIs im Vergleich zu Trizyklika durch Anderson [1] dar. Trotz der scheinbar recht inkonsistenten Befundlage über die verschiedenen Studien wurde ein schwacher Vorteil für die Trizyklika geschlussfolgert. Eine differenzierte Reanalyse dieses Studienmaterials 2 Jahre später durch dieselbe Gruppe führte die Metaanalysen getrennt nach verschiedenen Studiensettings durch [2]: So konnte festgestellt werden, dass jede der beiden Substanzklassen eine unterschiedliche relative Wirksamkeit bei ambulanten im Vergleich zu stationären Behandlungsbedingungen zeigte; Trizyklika waren unter stationären, SSRIs unter ambulanten Bedingungen überlegen; in die ursprüngliche Gesamtanalyse gingen mehr stationäre als ambulante Studien ein. Die Metaanalyse allein kann also keine Evidenz garantieren. Es ist vielmehr notwendig, Schichtungen in der Stichprobe zu entdecken, die zu differenten klinischen Schlussfolgerungen führen.

Argument 4

Das Resultat von Metaanalysen ist abhängig von der Art, wie die eingeschlossenen empirischen Untersuchungen gefunden werden. Das Ziel der Vollständigkeit allein reicht nicht aus; verschiedene systematische Suchstrategien nach relevanten empirischen Untersuchungen sind möglich und können zu unterschiedlichen Schlussfolgerungen führen.

Valide Metaanalysen setzen einen unverfälschten, ergebnisunabhängigen Zugriff auf alle zu einer Fragestellung informativen Studien voraus. Einerseits stellen Mehrfachpublikationen ein Problem dar, andererseits kann die Publikation von Studien von den Ergebnissen der Studien abhängen (nichtsignifikante Ergebnisse sind weniger – scheinbar – leicht interpretierbar und damit schwieriger zu publizieren als signifikante). Dies kann durch einen Funnel-Plot erkannt werden, ist aber nachträglich nicht korrigierbar. Ein solcher ergebnisabhängiger „Publikationsbias" konnte z. B. beim Vergleich neuerer Antipsychotika im Vergleich zu niedrigpotenten klassischen Antipsychotika festgestellt werden [16]. Es ist bemerkenswert, dass über diesen methodischen Mangel bei der gezogenen Schlussfolgerung hinweggegangen wird.

Melander et al. [17] haben das Problem „Publikationsbias" anhand von 5 zufällig ausgewählten Substanzen dargestellt; sie suchten erschöpfend nach publizierten wie auch nach noch nicht publizierten Studien zur Wirksamkeit jeder dieser Substanzen. Sie verglichen für jede dieser 5 Substanzen die durch Metaanalyse ermittelte globale Effektgröße (a) für alle verfügbaren Studien, publiziert oder nicht publiziert, (b) für alle publizierten Studien (ohne Berücksichtigung von überlappenden Stichproben), (c) für publizierte Studien nach Korrektur für Überlappungen zwischen den dargestellten Stichproben. Es resultierte eine erhebliche Variation der quantitativen globalen Effektgrößen, die zumindest für eine Substanz unter den verschiedenen Suchstrategien zur unterschiedlichen qualitativen Schlussfolgerung bezüglich der Wirksamkeit geführt hätte.

Die häufig als Referenzgrundlage gewählten Metaanalysen der Cochrane Collaboration oder von einzelnen Autoren wie Davis et al. [7], Geddes et al. [10, 11] und Leucht et al. [15, 16] beziehen sich ausschließlich auf publizierte Studien und unterliegen damit einer möglichen Verfälschung durch einen Publikationsbias. Eine mögliche Verfälschungsquelle kann allein schon aufgrund der Ergebnispräsentation in einzelnen Studien resultieren. Manchmal findet sich nämlich in den Publikationen einschlägiger Studien nicht genügend statistische Information zur Ermittlung der Effektstärke. Solche Studien müssen in Metaanalysen wegfallen (s. oben), die ja grundsätzlich auf Effektstärken zurückgreifen.

Aber selbst wenn systematisch auch Studien, die nicht publiziert wurden, in die Analyse aufgenommen werden, können sich noch erhebliche Diskrepanzen ergeben. Das zeigt eine Analyse über die Effekte von SSRIs auf die Anzahl von Suizidversuchen im Vergleich zu Plazebo aus jüngster Zeit. Hierzu wurden zeitlich parallel und unabhängig voneinander zwei Metaanalysen durchgeführt und gleichzeitig publiziert. Gunnell et al. [12] identifizierten alle plazebokontrollierten Studien in Medline und im Cochrane-Register, während Fergusson et al. [9] alle in der Registrierbehörde „Medicines and Health Care Production Regulation Agency" gemeldeten einschlägigen Studien zugrunde legten. Die Metaanalysen erbrachten für beide Suchstrategien unterschiedliche Resultate und qualitativ differente Schlussfolgerungen: Gunnell et al. fanden keine signifikante Differenz in Bezug auf die Häufigkeit von Suizidversuchen bei SSRIs im Vergleich zu Plazebo, während Fergusson et al. signifikant mehr Suizidversuche unter SSRIs berichteten.

Argument 5

Schlussfolgerungen von Metaanalysen hängen qualitativ von der Auswahl der Teststatistik ab, die zur Kombination der Ergebnisse von Einzelstudien benutzt wird.

In methodischer Hinsicht stellen Metaanalysen eine Gruppe von Verfahren zur statistischen Zusammenfassung quantitativer Untersuchungsergebnisse dar. Diese verschiedenen Verfahren sind untereinander austauschbar, ohne dabei jedoch hinlänglich robust zu sein: Verschiedene metaanalytische Verfahren kommen beim selben Datenbestand zu qualitativ unterschiedlichen Ergebnissen. Ein eindringliches Beispiel stellt der Vergleich von Metaanalysen zur Wirksamkeit unterschiedlicher Klassen von Antipsychotika dar: Leucht et al. [16] verglichen die klinische Wirksamkeit von Antipsychotika der zweiten Generation mit denen der ersten Generation. Die Autoren stellten eine überlegene Wirkung der ersteren fest und folgerten: Mögliche Vorteile in der Wirksamkeit von Antipsychotika der zweiten Generation sollten bei klinischen Therapieentscheidungen beachtet werden, diese sollten eher als konventionelle Antipsychotika genutzt werden. Diese Feststellung widerspricht den Resultaten einer Meta-analyse von Geddes et al. [10] zum selben Thema. Ein möglicher Erklärungsgrund könnte hierfür die umfangreichere Studienlage bei Leucht et al. sein. Daher griffen Geddes et al. [11] diesen Widerspruch auf und analysierten das Studienmaterial von Leucht et al. erneut und zwar unter Anwendung der von ihnen in einer früheren Publikation gewählten statistischen Analysetechnik. Überraschenderweise kamen sie aber nicht zu dem gleichlautenden Ergebnis wie Leucht et al.; es konnte erneut kein Wirksamkeitsunterschied zwischen beiden Substanzklassen festgestellt werden. Die kritische Differenz zwischen beiden sich widersprechenden Metaanalysen lag lediglich in einer unterschiedlichen Wahl der Teststatistik: Risiko-Differenz-Quotient vs. Log-odds-Verhältnis. Diese technische Differenz zwischen beiden Analysen desselben Materials führte also zu qualitativ unterschiedlichen Schlussfolgerungen. Damit muss die Validität der statischen Methode Metaanalyse angezweifelt werden.

Schlussfolgerung aus fünf Argumenten

Diese fünf Argumente plädieren dafür, Metaanalysen nicht automatisch als die höchste Ebene empirischer Evidenz anzusehen. Das Ziel, die Aussagen der einzelnen Therapiestudien vollständig in dem Ergebnis einer Metaanalyse aufgehen zu lassen, ist häufig nicht erreichbar. Auch die zugrunde liegenden statistischen Methoden sind so wenig robust, dass Me-

taanalysen keine Validität beanspruchen können.

Empirische Evidenz ist auch weiterhin primär durch geeignet geplante kontrollierte Studien zu erbringen. Diese sind adäquat zu planen, die Stichprobenumfänge sind aufgrund einer Power-Analyse zu quantifizieren, die teststatistischen Resultate sind im Rahmen der benutzten statistischen Theorie zu interpretieren. Absolute Sicherheit kann aus einem einzelnen Versuch nicht abgeleitet werden, aber Ergebnisse sind validierbar durch Replikationen. Die methodischen Grundlagen für dieses Vorgehen sind über ein Jahrhundert konsequent entwickelt worden und trugen zur zentralen Rolle der Biostatistik in der Medizin bei. Über diese Entscheidungslogik können auch Metaanalysen nicht hinausgehen, hierbei wird lediglich die quantitative Ergebnislage der Vergleichsstudien in einen quantitativen globalen Effekt zusammengefasst. Das Ausmaß bzw. die Sicherheit der zu einer Frage (Hypothese) vorliegenden Evidenz entscheidet sich an der Qualität und Anzahl der bekannten einschlägigen Studien entsprechend der Kriterien von Versuchsplanung und Teststatistik. Diese entscheidenden Gesichtspunkte werden durch die globale Effektgröße nicht reflektiert. Überbewertungen und Überinterpretationen von Metaanalysen wurden entsprechend auch in der klassischen biostatistischen Literatur sarkastisch und heftig kritisiert (z. B. [8]).

Welche Rolle können Metaanalysen in der Evidenzgewinnung spielen?

Die pauschale Anwendung von Metaanalysen zur Feststellung der relativen Wirksamkeit von Behandlungsmodalitäten ignoriert die Besonderheiten der eingeschlossenen Vergleichsstudien. Der wesentliche Nutzen differenzierter Metaanalysen ist dagegen die systematische Aufklärung der Quelle der Variation von Ergebnissen über die verschiedenen Studien zu einer einschlägigen Fragestellung. Das Auffinden solcher „Moderatorvariablen" in Sensitivitätsanalysen kann entweder hypothesenorientiert oder empirisch erfolgen. Für den letztgenannten Zweck sind Metaregressionsmodelle verfügbar, die klinisch relevante Einflussfaktoren feststellen können [20]. Anschließende gezielte hypothesenprüfende Studien können diese a posteriori gewonnenen Hypothesen explizit prüfen. So können mit Hilfe von Metaanalysen Differenzialindikatoren für spezifische Behandlungsverfahren abgeleitet werden (s. obiges Beispiel zum Vergleich SSRIs und Trizyklika).

Metaanalysen stellen also vorzugsweise Hilfsmittel für die Ableitung von Hypothesen dar, weniger aber die Methode für die Evidenzmaximierung für beliebige Therapieempfehlungen oder andere klinische Hypothesen. Für letzteres sind auch weiterhin adäquat und prospektiv geplante randomisierte kontrollierte Studien zur Prüfung der Wirksamkeit erforderlich und zur Ergebnisvalidierung eine hinlängliche Anzahl von Replikationen. Die summarische Wertung vorliegender empirischer Evidenz für eine Fragestellung kann die Besonderheiten und Limitationen, Schwächen und Stärken der einzelnen Studien nicht ignorieren.

Korrespondenzadresse

Prof. Dr. W. Maier
Klinik für Psychiatrie und Psychotherapie der Rheinischen Friedrich-Wilhelms-Universität
Sigmund-Freud-Straße 25, 53105 Bonn
Wolfgang.Maier@ukb.uni-bonn.de

Interessenkonflikt. Der korrespondierende Autor weist auf folgende Beziehungen hin:
Professor Maier hat von den folgenden Firmen Forschungsgelder erhalten, bzw. er ist Mitglied des Advisory Boards oder erhält Honorare für Vorträge: AstraZeneca, Böhringer, Bristol-Myers Squibb, Eli Lilly, Janssen Cilag, Lundbeck, Merck, Pfizer, Sanofi Aventis, Schering.
Von den folgenden Firmen hat Herr Professor Möller Forschungsgelder erhalten, ist Mitglied des Advisory Boards oder aber erhält Honorare für Vorträge: AstraZeneca, Bristol-Myers Squibb, Eli Lilly, Eisai, GlaxoSmithKline, Janssen Cilag, Lundbeck, Merck, Novartis, Organon, Pfizer, Sanofi Aventis, Sepracor, Servier, Wyeth.

Literatur

1. Anderson IM (1998) SSRIs versus tricyclic antidepressants in depressed inpatients: a meta-analysis of efficacy and tolerability. Depress Anxiety [Suppl 1] 7: 11–17
2. Anderson IM (2000) Selective serotonin reuptake inhibitors versus tricyclic antidepressants: a meta-analysis of efficacy and tolerability. J Affect Disord 58: 19–36
3. Antes G (1998) Evidence-based medicine. Internist 39: 899–908
4. Bauer M, Whybrow PC, Angst J et al. (Hrsg) (2004) Biologische Behandlung unipolarer depressiver Störungen. Behandlungsleitlinien der World Federation of Societies of Biological Psychiatry (WFSBP). Wissenschaftliche Verlagsgesellschaft, Stuttgart
5. Berger M, Antes G, Hecht H (2004) Die Lehrbuchrelevanz von evidenzbasierter Medizin und der Cochrane Collaboration. In: Berger M (Hrsg) Psychische Erkrankungen – Klinik und Therapie. 2. Aufl. Urban & Fischer, München, S 3–16
6. Bortz J, Döring N (Hrsg) (2002) Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 3. Aufl. Springer, Berlin
7. Davis JM, Chen N, Glick ID (2003) A meta-analysis of the efficacy of second-generation antipsychotics. Arch Gen Psychiatry 60: 553–564
8. Feinstein AR (1995) Meta-analysis: statistical alchemy for the 21st century. J Clin Epidemiol 48: 71–79
9. Fergusson D, Doucette S, Glass KC et al. (2005) Association between suicide attempts and selective serotonin reuptake inhibitors: systematic review of randomised controlled trials. Br Med J 330: 396–402
10. Geddes J, Freemantle N, Harrison P, Bebbington P (2000) Atypical antipsychotics in the treatment of schizophrenia: systematic overview and meta-regression analysis. Br Med J 321: 1371–1376
11. Geddes JR, Harrison P, Freemantle N (2003) New generation versus conventional antipsychotics. Lancet 362: 404
12. Gunnell D, Saperia J, Ashby D (2005) Selective serotonin reuptake inhibitors (SSRIs) and suicide in adults: meta-analysis of drug company data from placebo controlled, randomised controlled trials submitted to the MHRA's safety review. Br Med J 330: 385–389
13. Helmchen H (2002) Evidenz der Evidenz-basierten Medizin? Nervenarzt 73: 1–2
14. Jüni P, Altman DG, Egger M (2001) Assessing the quality of controlled clinical trials. Br Med J 323: 42–46
15. Leucht S, Barnes TR, Kissling W et al. (2003) Relapse prevention in schizophrenia with new-generation antipsychotics: a systematic review and exploratory meta-analysis of randomized, controlled trials. Am J Psychiatry 160: 1209–1222
16. Leucht S, Wahlbeck K, Hamann J, Kissling W (2003) New generation antipsychotics versus low-potency conventional antipsychotics: a systematic review and meta-analysis. Lancet 361: 1581–1589
17. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B (2003) Evidence b(i)ased medicine – selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. Br Med J 326: 1171–1173
18. Möller H-J, Maier W (2007) Probleme der „evidence-based medicine" in der Psychopharmakotherapie. Problematik der Evidenzgraduierung und der Evidenzbasierung komplexer klinischer Entscheidungsprozesse. Nervenarzt (im Druck)
19. Moncrieff J, Kirsch I (2005) Efficacy of antidepressants in adults. Br Med J 331: 155–157
20. Sterne JA, Gavaghan DJ, Egger M (2000) Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol 53: 1119–1129