Soz.- Präventivmed. 42 (1997) 95–104
0303-8408/97/020095-10 $ 1.50 + 0.20/0

Maria Blettner[1], Brigitte Schlehofer[1], Willi Sauerbrei[2]

[1] *Abteilung Epidemiologie, Deutsches Krebsforschungszentrum, Heidelberg*
[2] *Institut für Medizinische Biometrie, Universität Freiburg, Freiburg*

Grenzen von Metaanalysen aus publizierten Daten bei epidemiologischen Fragestellungen

Zusammenfassung

Metaanalysen aus epidemiologischen Studien nehmen an Bedeutung insbesondere dort zu, wo der Zusammenhang zwischen einem Risikofaktor und der Zielerkrankung kontroves diskutiert wird, bzw. wenn in Einzelstudien keine eindeutige Dosis-Wirkungsbeziehung gefunden wird. Ihre Ergebnisse werden somit für viele Fragestellungen aus dem Bereich Gesundheits-, Arbeits- und Umweltpolitik zur Beurteilung von Risiken genutzt. Häufig wird hierfür auf publizierte Daten zurückgegriffen. Es wird hier diskutiert, inwieweit diese Form von Metaanalysen dem gestellten Anspruch nach zuverlässigen Aussagen zur qualitativen und quantitativen Wirkung von Risikofaktoren gerecht werden kann. Insbesondere wird auf die Unterschiede hingewiesen, die zwischen Metaanalysen aus experimentellen Daten oder aus randomisierten klinischen Studien und denen aus epidemiologischen Studien bestehen. Es werden Argumente für Metaanalysen aus dem klinischen Bereich auf ihre Gültigkeit in Beobachtungsstudien untersucht, insbesondere werden die Probleme bei der Berechnung eines einzelnen gepoolten Schätzers aufgeführt. Die Argumente werden an zwei Beispielen aus der Literatur exemplarisch aufgezeigt.

Für viele Fragestellungen aus dem Bereich der Gesundheits-, Arbeits- und Umweltpolitik spielen Ergebnisse von epidemologischen Studien zur Beurteilung von Risiken eine entscheidende Rolle. Allerdings sind dabei präzise und quantitative Angaben gefordert, die in einer Grössenordnung liegen, die häufig die Möglichkeiten von Einzelstudien übersteigen. Um diesen Anforderungen gerecht zu werden, hat der Trend zu Metaanalysen in den letzten Jahren in vielen Bereichen der medizinischen Forschung, darunter auch der Epidemiologie, zugenommen. Metalanalysen sind quantitative Zusammenfassungen von Ergebnissen mehrerer Einzelstudien. Sie gehen über den klassischen Reviewartikel (qualitative Bewertung) hinaus, da sie den Anspruch haben, einen quantitativen und möglichst unverzerrten Schätzer, z.B. für das relative Risiko (RR) oder das Standardisierte Mortalitätsverhältnis (SMR), zu berechnen und Angabe über dessen Präzision (mittels eines Konfidenzintervalls) zu machen.

Metaanalysen werden vielfach dort durchgeführt, wo der Zusammenhang zwischen einem Risikofaktor und der Zielkrankheit kontrovers diskutiert wird und wo in Einzelstudien keine eindeutige (Dosis-Wirkungs-)Beziehung gefunden wurde. Häufig sind die einzelnen Studien zu klein, um zu klaren und statistisch stabilen Aussagen zu gelangen. In einer Metaanalyse soll dann die Konsistenz der einzelnen publizierten Ergebnisse untersucht werden. Darüber hinaus soll evaluiert werden, bis zu welchem Grad inkonsistente und heterogene Ergebnisse durch Zufallsschwankungen oder durch systematische Unterschiede zwischen den Studien erklärt werden können. Oft ist das Hauptziel der Metaanalyse, einen kombinierten und somit „verbesserten" Schätzer für den Effekt eines Risikofaktors anzugeben.

Epidemiologische Studien sind Beobachtungsstudien. Eine randomisierte Zuteilung der Exposition wie z.B. bei kontrollierten klinischen Studien oder bei toxikologischen Experimenten ist nicht möglich. Bei der Auswertung und Interpre-

tation der Ergebnisse ist deshalb der Einfluss von Störfaktoren (Confoundern) zu berücksichtigen. Wegen der Heterogenität der Einzelstudien und der fehlenden Randomisierung wird von einigen Autoren bestritten, dass eine Zusammenfassung von Ergebnissen aus Einzelstudien zu sinnvollen, aussagekräftigen Ergebnissen führt[1–3]. Selbst im Bereich klinischer Therapiestudien sind Metaanalysen aus publizierten Daten sehr umstritten[4], obwohl sich hinsichtlich dieser Beurteilung in den letzten Jahren ein Wandel andeutet. In einigen Arbeiten wird heute mehr Gewicht auf die Untersuchung der Heterogenität und nicht auf die Berechnung eines gemeinsamen Schätzers gelegt[5]. Trotz der kritischen Einschätzung werden im Bereich der Epidemiologie in den letzten Jahren viele Metaanalysen veröffentlicht; allein im Jahre 1994 finden sich in MEDLINE fast 100 Arbeiten mit den Schlüsselworten „Metaanalysen" und „Epidemiologie". Zusätzlich werden vielfach Metaanalysen als gesundheitspolitische Entscheidungshilfen durchgeführt, die nur in internen Berichten verwendet, aber nie veröffentlicht werden.

Wir diskutieren in dieser Arbeit, ob die von vielen Autoren bevorzugte „einfachste" Form der Metaanalyse, die aus einer quantitativen Zusammenfassung von *publizierten Ergebnissen* besteht, dem gestellten Anspruch nach zuverlässigen Aussagen zu qualitativen oder quantitativen Wirkungen von Risikofaktoren gerecht werden kann. Eine andere Art von Metaanalysen, basierend auf einer zusammenfassenden Auswertung der Originaldaten verschiedener Studien, auch gepoolte Auswertungen genannt, erfordert einen wesentlich grösseren Aufwand und wird allgemein als qualitativ besser angesehen. Möglichkeiten und Grenzen der gepoolten Auswertung sollen in einer späteren Arbeit getrennt untersucht werden.

Argumente und Bewertung

Es werden hier die wichtigsten Argumente für Metaanalysen, wie sie z.B. für den Bereich randomisierter klinischer Studien genannt werden (vgl. [6]), aufgeführt und für epidemiologische Studien bewertet. Diese Bewertung bezieht sich prinzipiell auf Metaanalysen aus publizierten epidemiologischen Studien und gilt nicht nur für qualitativ fragwürdige Publikationen. Wir setzen voraus, dass die Metaanalysen zumindest nach strikten Regeln und mit einem Studienprotokoll durchgeführt werden, in dem die wissenschaftliche Fragestellung, die genaue Beschreibung der Methoden, sowie die Kriterien, nach denen eine Zusammenfassung der Ergebnisse durchzuführen ist, festgelegt wird.

1. Argument
Metaanalysen ermöglichen es, einen einzelnen gemeinsamen quantitativen Schätzer für das relative Risiko* und dessen Konfidenzintervall anzugeben und damit eine quantitative Bewertung des Risikofaktors vorzunehmen

Bewertung:

Die Berechnung eines gemeinsamen Schätzers ist einer der wichtigsten Unterschiede zwischen einem klassischen Review und der Metaanalyse. Mit der Angabe dieses Parameters und dessen Konfidenzintervalls ist eine Bewertung gewünscht, die z.B. im Rahmen von Risikoabschätzungen oder gesundheitspolitischen Diskussionen eine Bedeutung hat. Da hier ein Hauptproblem der Metaanalysen liegt, werden mögliche Probleme der Berechnung eines einzigen gemeinsamen Schätzwertes aus verschiedenartigen Studien im nächsten Abschnitt ausführlich diskutiert.

2. Argument
Aufgrund der Vergrösserung der Fallzahl wird für den Schätzer des relativen Risikos eine genauere Präzision als in den Einzelstudien erwartet. Damit wird die statistische Macht für Hypothesentests vergrössert. Dies gilt insbesondere für die Untersuchung von seltenen Ereignissen (z.B. seltene Erkrankungen oder Subtypen von Tumoren), für die in Einzelstudien nicht genügend Fälle vorhanden sind

Bewertung:

Diese Aussage trifft im wesentlichen zwar für Metaanalysen von randomisierten Therapiestudien zu, nicht aber für Metaanalysen in der Epidemiologie. Die Vergrösserung der Fallzahl reduziert hier zwar den Zufallsfehler, nicht aber die Verzerrungen, die in Beobachtungsstudien eine erhebliche Rolle spielen können. Diese Verzerrungen ergeben sich zum Teil durch die fehlende Randomisation, weshalb eine Adjustierung für Confoundervariablen unbedingt zu fordern ist. Hinzu kommen Verzerrungen durch Selektionsbias oder Mess- und Erhebungsfehler. Die Vergrösserung des Stichprobenumfangs führt nicht notwendigerweise zu einer Verringerung dieser Verzerrung, die im allgemeinen unabhängig von der Stichprobengrösse ist. Es besteht die Gefahr, dass die Zusammenfassung von heterogenen Daten die Verzerrung der Schätzer sogar erhöhen kann.

3. Argument
In der Metaanalyse kann eine gleichartige Darstellung der Einzelergebnisse vorgenommen werden. Dies dient zur besseren Beurteilung der Ergebnisse der Einzelstudien und deren Heterogenität

Bewertung:

Die Untersuchung der Heterogenität ist eine zentrale Aufgabe von

* Es wird im folgenden immer von dem Relativen Risiko als Mass für den interessierenden Effekt gesprochen. Die Argumente gelten in gleicher Weise für andere Masse, wie das standardisierte Mortalitätsverhältnis (SMR) oder Inzidenzraten.

Metaanalysen bzw. jeglicher Art von Zusammenfassungen. Die in der Literatur vorgeschlagenen Methoden (siehe z.B.[5]) verlangen aber oft Informationen, die in Publikationen nicht angegeben sind. Ergebnisse von einzelnen Studien werden im allgemeinen sehr unterschiedlich dargestellt. Dieses Defizit kann in der Metaanalyse nicht behoben werden. Oft ist es auch nicht möglich zu entscheiden, ob Unterschiede im Ergebnis durch das Design der Studie, die Datenerhebung, statistische Auswertungsverfahren oder nur durch den Zufall erklärbar sind. Schon die unterschiedliche Bildung von Kategorien bei Risikovariablen oder die unterschiedliche Berücksichtigung (Adjustierung) von Confounders führt im allgemeinen zu einer Diskrepanz der Ergebnisse. Eine Beurteilung der Ursachen der Heterogenität kann daher aus publizierten Daten meist nur unzureichend vorgenommen werden.

4. Argument
Mit Hilfe einer Metaanalyse können in Einzelstudien generierte Hypothesen weitergehend untersucht werden. Es können Effekte für Subpopulationen (z.B. bestimmte Altersgruppen) untersucht werden, die bei der Analyse der einzelnen Studien nicht betrachtet werden, oder für die die Daten in den einzelnen Studien nicht ausreichend waren

Bewertung:

Im allgemeinen bedarf es einer erneuten Auswertung mit Originaldaten auf individuellem Niveau, um neue Hypothesen zu testen. Eine neue Kategorisierung der Risikovariablen oder die Berücksichtigung anderer Störvariablen ist aus publizierten Daten nicht machbar. Falls subgruppenspezifische Auswertungen publiziert wurden, sind diese meistens nicht einheitlich dargestellt (z.B. unterschiedliche Einteilung der Altersgruppen, verschiedene Einschlusskriterien für Tumorarten, keine getrennte Auswertung für Frauen und Männer). Metaanalysen aus publizierten Daten können hier also nicht zu schlüssigen Antworten führen. Verbesserungen können z.T. erreicht werden, wenn es gelingt, fehlende Angaben von den Autoren der Erstveröffentlichungen direkt zu erhalten.

5. Argument
Metaanalysen können durch die kritische Beurteilung vorhandener Studien zur Qualitätsverbesserung zukünftiger Studien beitragen

Bewertung:

Dies kann ein Gewinn von Metaanalysen, aber auch von qualitativen Reviews sein. Eine Auseinandersetzung mit der Qualität der Studien kann dazu führen, dass zukünftig methodische Standards definiert werden, um Einzelstudien zu verbessern. Werden die Einzelstudien kritisch evaluiert und verglichen, könnte dies zukünftige Studienleiter stimulieren, die Qualität des Designs, der Datenerhebung und der Auswertung zu verbessern und neue methodische Entwicklungen zu berücksichtigen.

6. Argument
Neue Forschungsfragen inhaltlicher sowie methodischer Art können durch die Metaanalyse aufgeworfen und präzisiert werden

Bewertung:

Dies kann sowohl eine Metaanalyse als auch ein qualitatives Review leisten. Beide können deutlich machen, welche Fragen trotz zahlreicher Einzelstudien zu einem Thema offen bleiben und weitere Forschung erfordern. Außerdem kann die in Metaanalysen beobachtete Abhängigkeit der Ergebnisse von der Studienform (z.B. systematische Unterschiede der relativen Risiken zwischen Fall-Kontrollstudien und Kohortenstudien) oder der Messinstrumente statistische und epidemiologische Methodenforschung stimulieren.

Probleme des gepoolten Schätzers

Eine wesentliche Erweiterung der Metaanalyse im Vergleich zum klassischen Review liegt in der Berechnung eines *gepoolten* Schätzers für den zu untersuchenden Effekt und dessen Konfidenzintervall. Dieser gemeinsame Effektschätzer ist das gewichtete Mittel der geschätzten Parameter der Einzelstudien, wobei die Gewichte in den meisten Fällen aus der Standardabweichung der Parameter in den Einzelstudien gebildet werden. In einigen Fällen wird auch eine Qualitätsbeurteilung der Studien berücksichtigt. Eine wesentliche Frage ist, ob die Ergebnisse aus den Einzelstudien zu einem einzigen Parameterschätzer sinnvollerweise zusammengefasst werden und zu einer interpretierbaren Ergänzung gegenüber der Darstellung aus einem qualitativen Review führen können. Einige Aspekte, die in vielen Bereichen eine Zusammenfassung der Ergebnisse erschweren oder sogar verbieten, sollen hier angeführt werden.

Heterogenität der Studiendesigns

In der Epidemiologie sind verschiedene Studiendesigns verbreitet, die je nach Fragestellung und Konstellation verschiedene Vor- und Nachteile aufweisen. Es ist fraglich, ob Daten aus verschiedenen Studiengruppen, z.B. Querschnittstudien, Kohortenstudien und Fall-Kontrollstudien, miteinander kombiniert werden sollten. Einige Autoren von Metaanalysen beschränken sich auf einen Studientyp, häufig findet sich aber auch eine gemeinsame Auswertung von Studien unterschiedlichen Designs (siehe Beispiel 1

und 2). Da das Design häufig in Bezug zum Thema gewählt wurde, ist eine Zusammenfassung der Ergebnisse, die aus verschiedenen Designs hervorgehen, nicht gerechtfertigt. Ausserdem ergibt sich durch diese Mischung eine zusätzliche Variabilität, die bei der Zusammenfassung nicht berücksichtigt werden kann.

Definition des Risikofaktors

Nur selten sind die untersuchten Risikofaktoren in verschiedenen Studien einheitlich definiert. Bei der Datenerhebung zur Umweltbelastung, zu Ernährungsgewohnheiten und anderer Lebensstilfaktoren werden unterschiedliche Messinstrumente (schriftlicher Fragebogen, persönliches Interview, Telefoninterview, direkte Messungen) verwendet, die zu unterschiedlichen Skalenniveaus führen. Selbst bei exakt definierten Faktoren, wie z.B. Gewicht, Anzahl der gerauchten Zigaretten oder berufliche Strahlendosis können unterschiedliche Kategorisierungen gewählt werden. Zudem ist die Definition der Referenzkategorie („nicht-exponierte" Personen) oft nicht einheitlich. Dies behindert eine sinnvolle Zusammenfassung der Daten oder erfordert die Reduzierung der Einflussgrösse zu einer binären Variablen mit den Ausprägungen „exponiert" oder „nicht-exponiert". Sinnvolle Untersuchungen zu Dosis-Wirkungsbeziehungen sind damit im allgemeinen nicht mehr möglich.

Definition der Zielkrankheit

Die Präzision der Diagnose der Zielkrankheit ist in den Studien teilweise unterschiedlich. So können Erkrankungen aufgrund rein klinischer Untersuchungen, Verdachtsdiagnosen, bildgebender Verfahren oder histologischer Befunde eingeschlossen werden. Unterschiedliche Studien nehmen auch eine unterschiedlich starke Differenzierung der Krankheit oder Krankheitsgruppen vor. Eine Zusammenfassung oder Aufgliederung in für die Metaanalyse relevante Diagnosegruppen ist aus publizierten Daten selten möglich und wird praktisch nie vorgenommen.

Definition der Studienpopulation

Die den Studien zugrundeliegenden Studienpopulationen unterscheiden sich häufig erheblich, da die Studien in unterschiedlichen Regionen, zu unterschiedlichen Zeitpunkten und mit unterschiedlichen Ausgangsfragestellungen durchgeführt werden. Je nach spezifischen Gegebenheiten werden Ein- und Ausschlusskriterien, z.B. bezüglich der Altersgrenzen, der Wohnorte oder der Zugehörigkeit zu nationalen, ethnischen oder sozialen Gruppen von den einzelnen Studien verschieden definiert.

Adjustierung für Confounder

Bei der Schätzung des relativen Risikos aus nicht-randomisierten Studien sind die Confoundervariablen adäquat zu berücksichtigen. Die hierfür üblichen Ansätze sind stratifizierte Analysen mit Berechnung des Mantel-Haenszel Schätzers oder die Festlegung eines logistischen Regressionsmodells. Zur Bewertung der Ergebnisse einer Einzelstudie ist die Berücksichtigung von relevanten Confoundern essentiell. Üblicherweise wird dazu eine ausführliche Modellbildung durchgeführt[7], wobei die Confounder einen starken Einfluss auf den Schätzer des relativen Risikos der interessierenden Einflussgrösse haben können. Für die Berechnung eines gepoolten Schätzers muss daher eine gewisse Homogenität der Einzelstudien bezüglich der Adjustierung für Confounder verlangt werden, bei publizierten Ergebnisse ist dies aber nicht der Fall. Die publizierten Schätzer stammen jeweils aus anderen Modellen und beschreiben damit einen anderen Effekt. Wenn z.B. in einer Fall-Kontrollstudie zur Untersuchung des Brustkrebsrisiko durch Ernährungsgewohnheiten für die Variable „Gesamtenergieaufnahme" adjustiert wird, so hat der Effekt der Variablen „Energieaufnahme durch Eiweiss" eine andere Bedeutung als in einem Modell, in dem für diese Variable nicht adjustiert wurde. Insbesondere können bei stark korrelierten Einflussgrössen einzelne Schätzer aus multivariaten Modellen nicht isoliert betrachtet werden. Hinzu kommt ein wichtiges statistisches Argument, das eine Kombination der Schätzer aus unterschiedlichen Modellen nicht zulässt: Eine Vergrösserung der Zahl der Variablen im Modell erhöht die Varianz des Schätzers für den interessierenden Risikofaktor, verringert aber im allgemeinen die Verzerrung. Da die Varianz im wesentlichen die Gewichtung des Schätzers bestimmt, erhält der Schätzer aus einer Studie ein geringeres Gewicht, wenn zusätzliche Faktoren zur Adjustierung benutzt werden, obwohl gerade dadurch eine Verringerung der Verzerrung erreicht werden sollte.

Gewichtungsfaktoren

Der bei einer Metaanalyse berechnete Schätzer für das relative Risiko ist ein gewichteter Mittelwert, wobei die Gewichtung umgekehrt proportional zur Varianz der Einzelschätzer ist, also im wesentlichen von der Grösse der Studien und der Modellwahl (z.B. der Anzahl der Confounder) abhängt. Das bedeutet aber, dass nur die durch den Zufall verursachten Variationen (random error), nicht aber Fehler oder Verzerrungen durch Selektionsmechanismen der Studienpopulation, durch ungenügende Berücksichtigung der Confounder oder Fehlklassifikation (systematic error) in die Gewichtung eingehen. Es wurde daher vorgeschlagen, die Qualität von Studien durch Bildung eines Scores

einzubringen, so daß „gute" Studien, die möglichst frei von Verzerrung und Bias sind, ein hohes Gewicht erhalten. Bisher wurden diverse Gewichtungssysteme vorgeschlagen[8,9], allerdings ist keines davon allgemein akzeptiert. Die Kritik an diesem Vorgehen[10] beruht unter anderem darauf, dass eine eindeutige Beurteilung der Qualität epidemiologischer Studien, z.B. eine Überlegenheit der Kohortenstudie im Vergleich zur Fall-Kontrollstudie, nicht möglich ist. Dies ist nicht vergleichbar zur Situation in der klinischen Therapieforschung, in der die randomisierte Therapiestudie als „beste" Studienform akzeptiert ist. Die Berücksichtigung der Qualität von Studien bei Metaanalysen wird aber auch hier sehr kontrovers diskutiert.

Modelle mit festen oder zufälligen Effekten

In den meisten Fällen wird zur Berechnung eines gepoolten Schätzers und seiner Variablen ein Modell mit festen Effekten (fixed effects model) angenommen. Dabei wird jede Studie als eine Stichprobe der gleichen Population angesehen, für die ein zugrundeliegender Effekt geschätzt wird. Dieser Ansatz geht lediglich von einer Variabilität innerhalb der Studie aus (within study variability). Unterschiede zwischen den Studien werden der zufälligen Stichprobenvariation zugeordnet. Als Alternative wird ein Modell mit zufälligen Effekten (random effects model) diskutiert, bei dem die Studien als zufällige Stichproben aus verschiedenen Populationen angesehen werden. Die Variabilität zwischen den Studien wird als ein Bestandteil der Gesamtvariabilität angesehen (between study variability) und daher als integraler Bestandteil auch im Modell berücksichtigt. Die wichtigste Implikation für eine Metaanalyse ist eine unterschiedliche Gewichtung der Studien, da sich bei Modellen mit zufälligen Effekten die Gewichte aus einer Kombination der „within" und „between study variabiality" bestimmen. Damit hat die Wahl des Modells einen grossen Einfluss auf Standardfehler und Konfidenzintervalle, die bei Modellen mit zufälligen Effekten im allgemeinen grösser sind. Bei einer grossen Variabilität zwischen den Studien kann das Konfidenzintervall beträchtlich grösser werden. Der gepoolte Schätzer des Expositionseffektes wird im allgemeinen nur gering beeinflusst. Wegen des grösseren Konfidenzintervalls sind die Resultate bei Modellen mit zufälligen Effekten somit seltener signifikant. Dies zeigen Berlin et al. (1989)[11], die die beiden Modellansätze anhand von 22 Metaanalysen randomisierter Studien vergleichen. Für weitere Unterschiede verweisen wir auf Dickersin & Berlin 1992[12].

Publikationsbias

In klinischen Studien wird der Publikationsbias von vielen Autoren als eine wesentliche Verzerrungsquelle für den gepoolten Schätzer angegeben. Im Gegensatz zu den klinischen Studien gibt es in der Epidemiologie kein Äquivalent zu „allen jemals ramdonisierten Personen zum Vergleich von Therapie A mit Therapie B". Die Gesamtpopulation aller Personen, über die Informationen zu einem Risikofaktor vorliegen, ist nicht definiert. So werden häufig in Fall-Kontrollstudien eine Reihe von Confoundervariablen erhoben, die aber nie getrennt ausgewertet oder veröffentlicht werden. Die Trennung zwischen Confoundervariablen und eigentlichen Risikofaktoren ist in Beobachtungsstudien nicht immer klar, und es stellt sich die Frage, ob auch solche Studien in eine Metaanalyse einbezogen werden sollten, bei denen für die Metaanalyse interessierende Variable nur als Confounder untersucht wurde. Es ist ausserdem offensichtlich, dass Studien mit signifikanten Ergebnissen häufiger veröffentlicht werden als „nicht signifikante" Studien. Hinzu kommt, dass oft aus der Vielzahl der erfassten Variablen – z.B. in sogenannten hypothesengenerierenden („fishing") Studien – nur eine Auswahl von „signifikanten Variablen" zur Veröffentlichung gelangt. Verzerrungen der gepoolten Schätzer durch einen solchen Publikationsbias sind in einem sogenannten *Funnelplot* (siehe Abbildung 1) zu erkennen. Diese Graphik wird für die Darstellung von univariaten Schätzern (z.B. Therapievergleich) in randomisierten Studien vorgeschlagen. In einem *Funnelplot* wird für jede Studie der Schätzer für den Effekt (z.B. relatives Risiko) gegenüber dem Umfang der Studie abgetragen. Bei Studien mit geringem Umfang ist die zufällige Streuung der Ergebnisse um den unbekannten wahren Wert grösser, während die Variation bei grossen Studien geringer ausfällt. Typischerweise sollte der *Funnelplot* daher eine Filterform aufweisen. Unter der Annahme, dass Studien mit geringem Stichprobenumfang bei einem negativen Ergebnis häufiger nicht veröffentlicht werden, kann die Filterform beim Vorliegen dieses Publikumsbias gestört sein. Allerdings ist diese Interpretation nur dann gerechtfertigt, wenn angenommen werden kann, dass alle Studien denselben Effekt schätzen. Ein *Funnelplot* kann bei epidemiologischen Studien nur dann zur Untersuchung des Publikationsbias eingesetzt werden, wenn die geschätzten Effekte der Einzelstudien nicht in unterschiedlicher Weise durch andere Faktoren (z.B. Confounder-Adjustierung) beeinträchtigt wurden. Die Evaluierung des *Funnelplots* kann daher nur als erster Schritt in der Analyse des Publikumsbias gesehen werden. Der Publikumsbias führt im all-

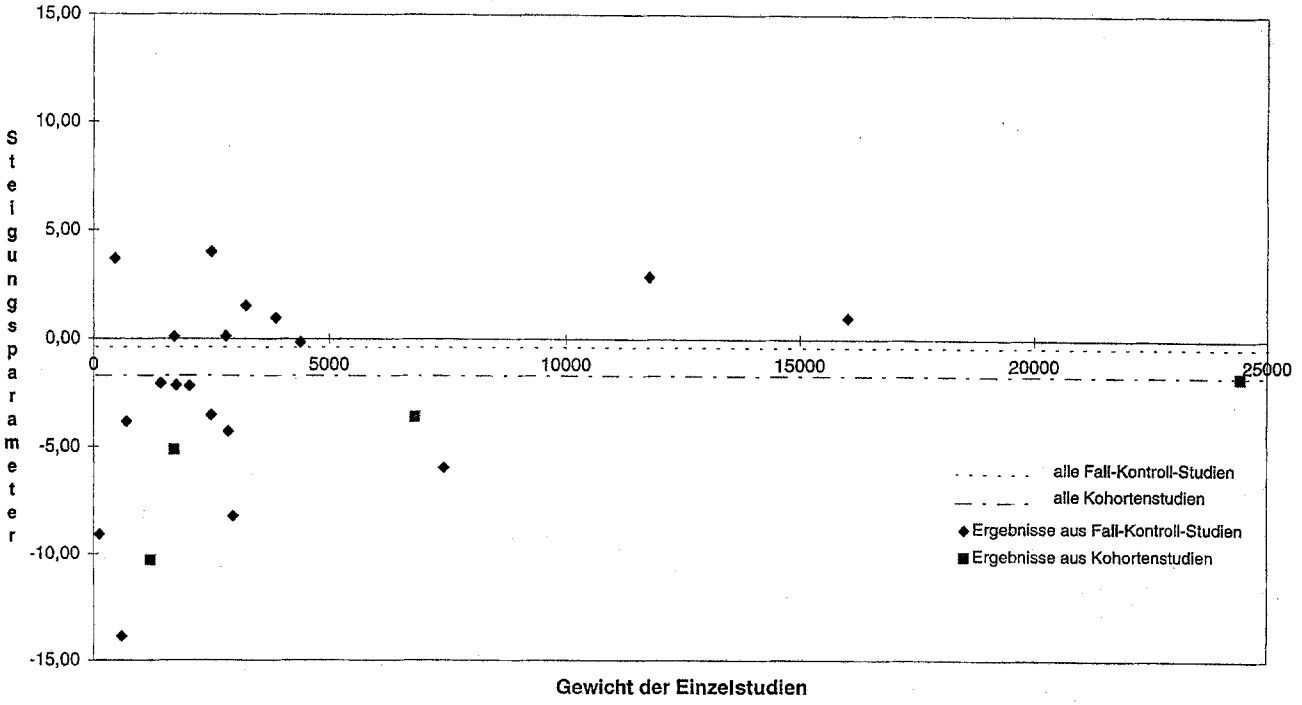

Abbildung 1. *Funnelplot: Darstellung der Ergebnisse der Einzelstudien nach Ursin et al.[14]: Brustkrebsrisiko für prämenopausale Frauen und Body-Mass-Index.*

gemeinen zu einer Überschätzung des Effektes, da die Studien, in denen keine „signifikanten" Ergebnisse gefunden wurden, seltener veröffentlicht werden.

Beispiele

Einige der genannten Probleme sollen anhand von zwei Beispielen publizierter Metaanalysen illustriert werden. Beide Arbeiten sind in international anerkannten Zeitschriften publiziert und beinhalten sowohl eine sorgfältige Literaturrecherche als auch eine ausführliche Diskussion vieler Probleme und Einschränkungen. Wir wollen hier nicht diese speziellen Beispiele kritisieren, sondern einige der oben genannten Punkte exemplarisch an zwei veröffentlichten Metaanalysen aus publizierten Daten verdeutlichen.

Beispiel 1

In der Arbeit von Washburn et al. (1994)[13] werden Ergebnisse von Studien zur Untersuchung des Zusammenhangs zwischen der Exposition gegenüber elektromagnetischen Feldern (EMF) und dem Auftreten von Krebs bei Kindern „meta-analysiert". Die Autoren nennen als Ziel ihrer Metaanalyse, zum Verständnis der Assoziation zwischen spezifischen Expositionen gegenüber EMF (Wohnsitz in der Nähe von Überlandleitungen, Transistoren, etc.) und bestimmten Erkrankungen (Leukämien, Lymphomen und Tumoren des Nervensystems) bei Kindern beitragen zu wollen. Nach einer gründlichen Literaturrecherche und der Datenabstraktion (relatives Risiko, Varianz, Anzahl der Confounder und andere relevante Informationen) aus den gefundenen Veröffentlichungen berechneten die Autoren ein gewichtetes Mittel der logarithmischen relativen Risiken, wobei die Gewichte proportional zur Varianz der Schätzer sind. Es wird ein random effects Modell verwandt, ohne diese Wahl näher zu begründen. Ein System für die Beurteilung der Qualität der Datenerhebung wird zwar beschrieben, bei der Berechnung der gepoolten Schätzer aber nicht berücksichtigt.

Probleme

– *Studientypen:* Insgesamt wurden 14 Studien in die Metaanalyse eingeschlossen, die bis Januar 1994 publiziert waren. Die Autoren machten keine Angaben über die Studientypen. Aus den Originalarbeiten ist zu sehen, dass sowohl Fall-Kontrollstudien als auch Kohortenstudien eingeschlossen wurden.

– *Definition des Risikofaktors:* Die Exposition gegenüber EMF wurde in den Studien sehr unterschiedlich ermittelt: Vier Studien erfassten eine qualitative Einschätzung der Exposition (wire code), in einer Studie wurden Messungen, in zwei Studien Berechnungen der elektromagnetischen Feldstärke vorgenommen, in den anderen Studien wurden nur Entfernungen zu diversen Expositionsquellen be-

trachtet (z.B. Hochspannungsleitungen, elektrische Verteilereinrichtungen).
– *Definition der Confoundervariablen:* Die meisten angegebenen relativen Risiken in den Originalarbeiten wurden nach Alter (10mal), Geschlecht (8mal), und/oder zusätzlich nach anderen Faktoren adjustiert (10mal), einmal wurde nicht für Störvariablen adjustiert. In einem Fall ist es den Autoren nicht gelungen, die Adjustierungsfaktoren zu ermitteln.
– *Definition der Zielkrankheit:* In 13 Studien wurde „Leukämie bei Kindern" untersucht. Es werden keine Aussagen gemacht, ob in den Einzelstudien alle Typen von Leukämien zusammengefasst werden oder ob nur bestimmte Leukämieformen untersucht wurden.
– *Studienpopulation:* Neben den grossen regionalen Unterschieden der betrachteten Studiengruppen (Taiwan, Mexiko, England, USA, Nordeuropa, Australien und Griechenland), variert die obere Altersbegrenzung der Fälle zwischen 11 und 21 Jahren, so dass zum Teil nur Kinder, zum Teil auch Jugendliche eingeschlossen wurden.
– *Ergebnis:* Trotz der qualitativ ersichtlichen und quantitativ ermittelten Heterogenität ($p = 0.02$) geben Washburn et al.[12] eingepooltes relatives Risiko und das zugehörige Konfidenzintervall an (RR = 1,49; 95% KI = 1,1–2,0). Die Herausnahme einer Studie, die zu sehr abweichenden Ergebnissen geführt hatten, verändert diesen Schätzwert beträchtlich (RR = 1,34; 95% KI = 1,30–1,74) und verringert auch die beobachtete Heterogenität. Die relativen Risiken der Einzelstudie variieren zwischen 0,5 und 6,0, wobei beide Extremwerte aus schwedischen Studien stammen. Die neueste und grösste Studie zeigt eine nichtsignifikante Erhöhung des relativen Risikos (RR = 1,2)*.

Bewertung

Die Einzelstudien sind untereinander sehr heterogen, insbesondere bezüglich der Erfassung der Exposition und der Betrachtung von Confoundervariablen. Eine ausführliche Diskussion der Unterschiede zwischen den Einzelstudien fehlt. Der gepoolte Schätzer wird deutlich von der neuesten und grössten Studie bestimmt, was zwar zu seiner Glaubwürdigkeit beiträgt, vielleicht aber auch zeigt, dass die Metaanalyse keine neuen Erkenntnisse über diese grosse Einzelstudie hinaus erbringt.

Beispiel 2

1995 führten Ursin et al.[14] eine Metaanalyse durch, um den Einfluss des Body-Mass-Index (BMI) auf die Entstehung von prämenopausalem Brustkrebs zu quantifizieren. Die Autoren geben an, dass eine Erhöhung des BMI zwar mit einem vergrösserten Risiko an postmenopausalem Brustkrebs zu erkranken assoziiert ist, dass aber trotz verschiedener Studien zu dieser Problematik die Beziehung zwischen Grösse und Gewicht und dem Brustkrebsrisiko vor der Menopause ungeklärt bleibt.

Probleme

– *Studientypen:* Insgesamt wurden 23 Studien, davon 19 Fall-Kontrollstudien (5 mit Krankenhauskontrollen und 14 mit Populationskontrollen) und 4 Kohortenstudien in die Metaanalyse aufgenommen.
– *Definition des Risikofaktors:* In allen Studien wurde die Häufigkeitsverteilung für BMI angegeben, wobei sowohl die Zahl der Kategorien als auch die Schnittstellen erheblich schwanken. Die Autoren der Metaanalyse betrachteten dann entweder den Median oder den Mittelwert für die einzelnen Kategorien als Score für die Untersuchung des Risikos in Abhängigkeit vom „BMI". 16 Fall-Kontrollstudien ermitteln Grösse und Gewicht zum Zeitpunkt des Interviews, in drei Fall-Kontrollstudien und in allen Kohortenstudien wurde das Gewicht von einem früheren Zeitpunkt verwendet. Eine Umrechnung oder Anpassung dieser Werte wurde nicht vorgenommen. In einigen der einbezogenen Studien wurde BMI als einer der interessierenden Risikofaktoren betrachtet, während in einigen anderen Studien BMI nur als Confounder betrachtet wurde, beispielsweise dann, wenn Ernährungsfaktoren untersucht wurden.
– *Definition der Confoundervariablen:* Es wurde aus jeder Publikation das relative Risiko aus dem umfangreichsten Modell ausgewählt. Ursin et al. geben eine Liste von Confoundervariablen an, die sie selbst für adäquat einschätzen. Allerdings wurden nur in sieben Fall-Kontrollstudien und in drei Kohortenstudien (!) eine oder mehrere dieser Variablen berücksichtigt.
– *Definition der Zielkrankheit:* In allen Studien werden Brustkrebsfälle bei prämenopausalen Frauen untersucht, die Definition der Krankheit kann daher als einheitlich angesehen werden.
– *Studienpopulation:* 21 Studien stammen aus westlichen Ländern (Europa und Nordamerika), jeweils eine Studie aus Japan und aus Israel, so dass Länder mit sehr unterschiedlichen Inzidenzraten und sehr unterschiedlicher Verteilung der Risiko- und Confounderfaktoren einbezogen wurden. Die Definition der eingeschlossenen Personen als „prämenopausal" geschah in den Einzelstudien zum Teil über das Alter (6mal), nach dem Menopausenstatus (16mal) oder nach sonstigen Kriterien (2mal); 6 Studien schliessen Frauen mit Hysterektomie aus.

* Alle hier genannten Schätzer und Konfidenzintervalle werden nur aus der Metaanalyse [12] zitiert. Die Originalarbeiten wurden zwar verglichen, aber auf Unstimmigkeiten soll hier nicht eingegangen werden.

– *Ergebnisse:* Die Autoren berechneten ein Regressionsmodell, in dem ein Steigerungsparameter für das relative Risiko als Funktion des BMI berechnet wird. Diese Angaben werden dann umgerechnet, so dass das relative Risiko (bezogen auf 8 Einheiten des BMI) angegeben wird. Für die vier Kohortenstudien und die 19 Fall-Kontrollstudien wurden getrennt Auswertungen vorgenommen, wobei jeweils ein Schätzer aus dem fixed-effect-model und dem random-effect-model angegeben wurde. Auf die Angabe eines gemeinsamen Schätzers aus allen 23 Studien wurde wegen der grossen Heterogenität verzichtet (Heterogenitätstest: $p = 10^{-8}$), allerdings sind auch die Ergebnisse innerhalb jedes Studientyps untereinander sehr heterogen. Abbildung 1 zeigt die Ergebnisse der einzelnen Studien in Form eines Funnelplots (entnommen aus Tabelle 1 in [14]). Diese Darstellung legt die Vermutung nahe, dass von einer selektiven Veröffentlichung von kleinen Studien ausgegangen werden muss.

Bewertung

Auch innerhalb des gleichen Studientyps zeigt sich eine auffallend hohe Heterogenität der Ergebnisse (siehe Tabelle 1), insbesondere variieren die Ergebnisse der Fall-Kontrollstudien mit kleinen Fallzahlen erheblich. Auffallend ist auch, dass das gepoolte relative Risiko aus den Kohortenstudien ausserhalb des Konfidenzintervalls der Fall-Kontrollstudien liegt. Die Ergebnisse werden insbesondere von einer grossen norwegischen

Merkmal	**Beispiel 1 (Washburn et al., 1994)**	**Beispiel 2 (Ursin et al., 1995)**
Anzahl der gefundenen Studien	Keine Angabe	27 oder 30
Anzahl der eingeschl. Studien	13 (für Leukämie)	23
Fall-Kontrollstudien	Keine Angabe[1]	19
Kohortenstudien	Keine Angabe[1]	4
Umfang der Studien	Anzahl der Leukämiefälle	„Gewicht"[2]
grösste Studie	832	24414
kleinste Studie	3	132
Adjustierung für Confounder (ausser Alter)		
RR nicht adjustiert	2	13
1 Confounder	2	
2 Confounder	6	10
Mehr als 3 Confounder	2	
Keine Angaben	1	
Gepooltes RR	RR 95 % KI	RR 95 % KI
Für alle Studien gemeinsam	1,49 (1,11–2,00)[3]	Keine Angabe
Für Kohortenstudien	Keine Angabe	0,81 (0,75–0,88)[4]
		0,70 (0,54–0,91)[3]
Für Fall-Kontrollstudien	Keine Angabe	0,95 (0,90–1,01)[4]
		0,88 (0,76–1,02)[3]
RR aus Einzelstudien	RR	RR[5]
RR aus kleinster Studie	6,00	0,48
RR aus grösster Studie	1,30	0,87
Minimales RR	0,20	0,33
Maximaler RR	6,00	1,38

[1] Nach unseren Recherchen: 1 Kohortenstudie, 1 Pilotstudie für Fall-Kontrollstudie, 11 Fall-Kontrollstudien.
[2] Keine Angaben zur Grösse vorhanden.
[3] Random-effects-model.
[4] Fixed-effects-model.
[5] Verringerung pro 8 Einheiten des BMI.

Tabelle 1. *Zusammenfassung der Metaanalysen von Washburn et al.[13] und Ursin et al.[14].*

Kohortenstudie dominiert, die 74% des Gewichts der Kohortenstudien ausmacht, während die Fall-Kontrollstudien von einer internationalen, multizentrischen und einer US-amerikanischen Studie dominiert werden (zusammen 47% der Daten der Fall-Kontrollstudien). Insgesamt kann der Interpretation der Autoren, dass Übergewicht gegen prämenopausalen Brustkrebs schützt, zugestimmt werden. Eine quantitative Bewertung lässt die beobachtete Heterogenität allerdings nicht zu.

Zusammenfassende Beurteilung zu Metaanalysen aus publizierten Daten

Unsere Hauptkritik an Metaanalysen aus publizierten Daten bezieht sich auf die Berechnung eines gemeinsamen Effektschätzers und dessen Konfidenzintervall. Wir sind der Meinung, dass Metaanalysen aus publizierten epidemiologischen Studien wegen der Heterogenität der Studiendesigns und insbesondere wegen der unterschiedlichen Adjustierung der publizierten Schätzer nicht geeignet sind, einen präzisen und unverzerrten gemeinsamen Effektschätzer für Risiken zu berechnen. Sie können auch nur bedingt zur Untersuchung der Heterogenität der Einzelstudien genutzt werden, da die dazu notwendige Information in den Publikationen meistens fehlt. Falls es gelingt, von den Autoren zusätzliche Informationen zu bekommen, ist hier sicherlich eine Verbesserung möglich. Für die Beurteilung der Qualität von Einzelstudien können solche Metaanalysen jedoch einen wichtigen Beitrag leisten, der zu Qualitätsverbesserung zukünftiger Studien und zu Standardisierungen führen kann.

Erkenntnisse aus Metaanalysen von publizierten Daten gehen nur dann über gute Reviews hinaus, wenn ein striktes Protokoll für den Einschluss und die Beurteilung von Studien eingehalten wird. Nur wenn die einbezogenen Studien ausreichend homogen in Bezug auf Design, Datenerhebung und Confounderadjustierung sind, ist es sinnvoll, einen gemeinsamen Effektschätzer zu berechnen. Bei vorhandener Heterogenität zwischen den Studien sind deren Ursachen zu untersuchen, auf einen gemeinsamen Schätzer sollte dann jedoch verzichtet werden. Werden trotz vorhandener Heterogenität zwischen den Studien in Metaanalysen gepoolte Effektschätzer angegeben, so wird die Illusion einer Genauigkeit und Gültigkeit vorgetäuscht, die nicht vorhanden ist. Dieses Vorgehen trägt viel zum Misskredit von Metaanalysen bei. Die Bewertung der Stärke eines Risikofaktors kann und darf nicht auf solche schwachen Argumente gestellt werden, insbesondere dann nicht, wenn aus solchen Metaanalysen Schlüsse gezogen werden, die z.B. zu gesundheitspolitischen oder arbeitsrechtlichen Konsequenzen führen.

Metaanalysen aus publizierten Daten epidemiologischer Studien sind allenfalls dann sinnvoll, wenn nur Studien, die methodologisch korrekt durchgeführt wurden und über gleichartige Designs und Auswertungen verfügen, in der Analyse berücksichtigt werden. Diese Einschränkungen dürfen aber auch nicht dazu führen, dass ein dadurch entstehender Selektions- bzw. Publikumsbias den Schätzer des Effekts verzerrt. Es ist zu hoffen, dass mit der Verbesserung des Standards in der epidemiologischen Forschung und der Vereinheitlichung der Methodik auch die Methodik von Metaanalysen verbessert werden kann. Es ist in jedem Fall zu empfehlen, zusätzliche Informationen von den Autoren der Originalarbeiten einzuholen. Wie bei randomisierten Studien in der klinischen Forschung werden auch in der Epidemiologie mehr und mehr Metaanalysen mit Hilfe der Originaldaten auf individueller Basis durchgeführt. Dieser Ansatz bietet Möglichkeiten, einige der genannten Schwierigkeiten zu überwinden und kann zu qualitativ zuverlässigeren Analysen führen.

Summary

Limits of metaanalysis from published data for epidemiological research

Metaanalyses of epidemiological studies have increased during the last years and are often used to evaluate the effect of risk factors which are inconsistent in different studies, mainly for small risk factors. Very often a metaanalysis is performed from published data. In this article we discuss this form of a metaanalysis and investigate whether the requirement to get reliable information is achievable with it. We mainly ask questions whether qualitative and quantitative dose-response analysis can be performed. We point out the differences between metaanalysis from experimental data and clinical randomized studies and epidemiological studies. We discuss different arguments that were given for performing metaanalysis in clinical trials and investigate whether they are also valid in observational studies. We mainly concentrate on the problem of estimating a single pooled risk estimate. Two examples from literature are used to show problems with metaanalysis from published data.

Résumé

Limites de la métaanalyse de données publiées dans la recherche épidémiologique

L'importance de la métaanalyse d'études épidémiologiques est accrue au cours des dernières années, surtout dans le cas de résultats très différents d'études ayant le même objet, par exemple quand l'effet d'un certain facteur à risque n'est pas clairement démontré. Les résultats de ces études sont largement utilisées pour évaluer les risques dans les secteurs de la santé et de la politique de l'emploi et de l'environnement. Les métaanalyses sont souvent basées sur des données publiées. Cet article discute dans quelle mesure cette forme de métaanalyse peut livrer des résultats fiables en ce qui concerne l'effet qualitatif et quantitatif de facteurs à risque. Il met en évidence les différences entre une étude épidémiologique et la métaanalyse de données expérimentales et d'études cliniques randomisées. La validité des arguments donnés pour la métaanalyse d'études cliniques y est discutée pour le cas d'études d'observation, avec une attention particulière au problème d'estimer un seul estimateur combiné. Deux exemples tirés de la littérature illustrent la discussion.

Literaturverzeichnis

1 *Spitzer, WO.* Meta-meta-analysis: unanswered questions about aggregating data. J Clin Epidemiol 1991; *44*:103–107.

2 *Feinstein AR.* Meta-Analysis: Statistical alchemy for the 21st century. J Clin Epidemiol 1995; *48*:71–79.

3 *Shapiro S.* Meta-Analysis/Shmeta-analysis. Am J Epidemiol 1994; *140*:771–778.

4 *Thompson SG, Pocock SJ.* Can meta-analyses be trusted? Lancet 1991; *338*:1127–1230.

5 *Thompson SG.* Why sources of heterogeneity in meta-analysis should be investigated. Br Med J 1994; *309*:1351–1355.

6 *Fleiss JL, Gross AJ.* Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. J Clin Epidemiol 1991; *44*:127–139.

7 *Blettner M, Sauerbrei W.* Influence of model-building strategies on the results of a case-control study. Stat Med 1993; *12*:1325–1338.

8 *Friedenreich CM, Brant RF, Riboli E.* Influence of methodologic factors in a pooled analysis of 13 case-control studies of colorectal cancer and dietary fiber. Epidemiology 1994; *5*:66–79.

9 *Longnecker MP, Berlin JA, Orza MJ, Chalmers TC.* A meta-analysis of alcohol consumption in relation to risk of breast cancer. JAMA 1988; *260*:652–656.

10 *Greenland S.* Quality scores are useless and potentially misleading. Reply to "Re: A critical look at some popular meta-analytic methods." Am J Epidemiol 1994; *140*: 300–301.

11 *Berlin JA, Laird NM, Sacks HS, Chalmers TC.* A comparison of statistical methods for combining event rates from clinical trials. Stat Med 1989; *8*:141–151.

12 *Dickersin K, Berlin JA.* Meta-analysis: State-of-the-science. Epid Reviews 1992; *14*:154–176.

13 *Washburn EP, Orza MJ, Berlin JA, Nicholson WJ, Todd AC, Frumkin H, Chalmers TC.* Residential proximity to electricity transmission and distribution equipment and risk of childhood leukemia, childhood lymphoma, and childhood nervous system tumors: systematic review, evaluation, and meta-analysis. Cancer Causes and Control 1994; *5*:299–309.

14 *Ursin G, Longnecker MP, Haile RW, Greenland S.* A meta-analysis of body mass index and risk of premenopausal breast cancer. Epidemiology 1955; *6*:137–141.

Korrespondenzadresse

Priv. Doz. Maria Blettner
Abteilung Epidemiologie
Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 280
Postfach 10 19 49
D-69120 Heidelberg