

Methoden zur Aggregation mehrerer Studien

Wie lassen sich Studien an mehreren Zentren (Studienorten) und durch verschiedene Teams durchführen? Wie können Ergebnisse vieler, individuell geplanter Forschungsarbeiten durch Reviews oder Meta-Analysen zusammengefasst werden?

Bei der Durchführung einer einzelnen Studie kommt meistens nur eine Messmethode zum Einsatz und es gibt oft keine ausreichend großen und heterogenen Stichproben, die die Generalisierung der Ergebnisse auf die gewünschte Population erlauben würden. Die Generalisierbarkeit kann erhöht werden, indem eine Vielzahl von Studienvariablen variiert wird. Ergebnisse gelten als valider, wenn nicht nur ein Forscher oder Forscherteam ein Untersuchungsergebnis replizieren kann, sondern wenn auch andere Forschende zu vergleichbaren Resultaten kommen. Ähnliches gilt auch, wenn die Befunde auf andere Stichproben repliziert (wiederholt) werden können. Falls im Gegensatz hierzu zum Beispiel ein bestimmter Zusammenhang nur bei Freiburger Psychologiestudierenden gefunden wird, muss der Frage nachgegangen werden, ob und wie sich Freiburger Studierende von anderen Studierenden bedeutsam unterscheiden und was das für die Verallgemeinerbarkeit des Ergebnisses bedeutet.

1 Planung multizentrischer Studien

Versuche zur Erlangung der Generalisierbarkeit können im Vorfeld geplant oder auf Grundlage schon durchgeführter Studien umgesetzt werden (Forschungsprogramme oder Meta-Analysen). Bei einem Forschungsprogramm oder einer multizentrischen Studie erfolgt, oft durch einen Förderer (z. B. Deutsche Forschungsgemeinschaft; DFG), eine bundesweite Ausschreibung. Durch Ankündigung eines Programms sollen möglichst viele Forschende erreicht werden. Interessierte Arbeitsgruppen reichen in einer ersten Ausschreibungsrunde eine kurze Projektskizze ein. Wird sie angenommen, so kann sie durch einen längeren Antrag erweitert werden.

Die Fördernden haben bei dieser Form der Ausschreibung die Möglichkeit, Forschende mit ähnlichen Studien und Ideen zusammenzubringen, sodass **Synergieeffekte** genutzt werden können. Haben sich die Förderer entschieden, welche Projekte sie mit finanziellen Mitteln unterstützen wollen, erfolgt ein Controlling der Projekte durch die Abgabe von Zwischen- und Endberichten.

Teilweise werden auch **Arbeitsgruppen** zu bestimmten Themen gegründet, die in einem intensiven Austausch stehen. Der Erfolg eines Projekts wird anhand der Qualität des Abschlussberichts und der Publikationen der Projektergebnisse gewertet. Ein großer Vorteil dieses Vorgehens ist, dass bestimmte Fragestellungen von Experten aus unterschiedlichen Professionen bearbeitet werden können. Beispielsweise kann ein Forschungsprogramm zur Erfolgsmessung bei Suchtbehandlung Teilprojekte beinhalten, die unabhängig voneinander rechtliche, sozioökonomische, psychosoziale oder medizinische Aspekte untersuchen.

2 Narrative Reviews und Meta-Analysen

Narrative Reviews und Meta-Analysen befassen sich mit der Zusammenfassung von Studienergebnissen verschiedener Studien zu einer Fragestellung – es sind Überblicksarbeiten. Warum aber werden Studienergebnisse überhaupt zusammengefasst? Für wen ist das wann sinnvoll?

Im Moment wird eine exponentiell anwachsende Anzahl von sog. **Primärstudien** publiziert. Hierbei handelt es sich um Arbeiten mit den Ergebnissen aus neuen Untersuchungen und Experimenten. Die Anzahl dieser Publikationen ist so groß, dass es unmöglich ist, diese „tägliche Flut“ von Detailinformationen zu überschauen und zu rezipieren. Allein im Fachbereich Medizin werden in über 10.000 Fachzeitschriften weltweit jährlich über zwei Millionen Zeitschriftenartikel veröffentlicht (www.cochrane.de/deutsch/ccbackg.htm). Dies bedeutet, dass es im Mittel 5.500 neue Artikel pro Tag, bzw. vier neue Publikationen pro Minute gibt. Selbstverständlich sind für Forschende oder Praktizierende nicht alle Arbeiten von gleicher Relevanz – die Relevanz kann allerdings oft erst nach dem Lesen einer Arbeit bewertet werden.

Reviews oder Meta-Analysen als **systematische Übersichtsarbeiten** ermöglichen, sich gezielt über den Forschungsstand zu bestimmten Fragestellungen zu informieren. Erst der Überblick ermöglicht, psychologische oder medizinische Entscheidungen in der täglichen Anwendung im Sinne einer „evidence based

medicine“ (EBM, evidenzbasiert, auf bestätigten Erfahrungen beruhend) stets anhand des aktuellen Wissensstandes zu treffen.

Aber auch in der **Grundlagenforschung** ist eine Zusammenfassung von wissenschaftlichen Ergebnissen zur Gewinnung eines Überblicks bei bestimmten Forschungsfragestellungen sinnvoll. Somit werden Reviews und Meta-Analysen den Anforderungen der folgenden **Zielgruppen** gerecht:

- Praktiker: Da seit Studium und Ausbildung oft viele Jahre / Jahrzehnte vergangen sind, kann so das vorhandene Wissen aktualisiert und erweitert werden.
- Lehrende / Studierende: Meta-Analysen geben einen Überblick, der meistens aktueller ist als das Wissen aus Lehrbüchern.
- Forschende: Überblicksarbeiten fassen den aktuellen „Gesamtforschungsstand“ in einem Bereich zusammen, sodass die Bildung neuer Hypothesen angeregt wird.

2.1 Prototypen der Analyse

Seitdem in den siebziger Jahren des letzten Jahrhunderts erstmalig der Begriff Meta-Analyse (meta analysis) für die quantitative Zusammenfassung einzelner Untersuchungsergebnisse verwendet wurde (Glass 1976), ist dieser Begriff immer häufiger in Publikationen zu finden, was belegt, dass diese Methodik vermehrt angewendet wird. Es haben sich eine Vielzahl von differierenden Methoden zur Meta-Analyse von Ergebnissen entwickelt, die Blettner, Sauerbrei, Schlehofer, Scheuchenpflug und Friedenreich (1999) zu den folgenden vier Prototypen zusammengefasst haben:

- Typ I: Reviews als Zusammenfassung von Studien, wobei die Studienergebnisse primär deskriptiv beschrieben werden.
- Typ II: Meta-Analysen, die Daten aus publizierten Studien verwenden. Dies ist die gebräuchlichste Form.
- Typ III: Auf Basis der individuellen Rohdaten aller integrierten Studien werden statistische Kennwerte ermittelt (retrospektives Vorgehen).
- Typ IV: Erfassung der individuellen Rohdaten aller integrierten Studien, wobei in diesem Fall die Studien noch nicht durchgeführt wurden, sondern in naher Zukunft geplant sind (prospektives Vorgehen). Auch hier werden die statistischen Kennwerte auf Basis der Rohdaten berechnet.

2.2 Reviews (Typ I)

Es erfolgt eine Beschreibung der verfügbaren Literatur zu einer Fragestellung. Die Ergebnisse jeder Studie werden kurz beschrieben, aber nicht zu einem Ergebnis zusammengefasst (im Gegensatz zu Meta-Analysen vom Typ II bis IV). Teilweise werden mehrere hundert Studien zu einer Fragestellung „nur“ deskriptiv dargestellt. Im Allgemeinen erfolgt diese Bewertung durch mehrere Reviewer, wobei es sich um eine eher „grobe“ Einteilung in zwei beziehungsweise drei Kategorien handelt. Studien können folgendermaßen bewertet werden:

- + (positiv): In diese Kategorie fallen Studien, deren Ergebnisse signifikante Unterschiede in die erwartete Richtung zeigen. Eine Psychotherapieerfolgsstudie belegt z. B., dass die subjektive Befindlichkeit der Patienten sich verbessert hat.
- = (kein Unterschied): Hier werden Studien eingeordnet, die keine signifikanten Unterschiede zwischen Gruppen gefunden haben. Die Ergebnisse der Psychotherapieerfolgsstudie weisen weder auf eine Verbesserung noch auf eine Verschlechterung der subjektiven Befindlichkeit durch Therapie hin.
- - (negativ): Die Ergebnisse der Studien dieser Kategorie zeigen signifikante Unterschiede entgegen der erwarteten Richtung. In der Meta-Analyse werden z. B. Studien erfasst, die einen negativen Therapieeffekt belegen. Die Patienten würden sich nach der Therapie im Mittel subjektiv schlechter fühlen.

Die statistische **Auswertung** dieser Bewertungen erfolgt über einen sog. *Box Score* oder ein *Vote Counting*. Gezählt wird, wie viele Studien den drei Kategorien (positiv / neutral / negativ) angehören. Teilweise werden die negative und die neutrale Kategorie einer Kategorie zusammengefasst (positiv / neutral und negativ). Über die Kategoriehäufigkeiten kann per χ^2 -Test ermittelt werden, ob Studien der einen Kategorie signifikant häufiger sind als Studien, die in eine andere Kategorie fallen. Die Interpretation dieser Box Scores und die Suche nach Gemeinsamkeiten bei „positiven“ und „negativen“ Studien können Hinweise auf potenzielle Moderatorvariablen geben.

Vorteil des Verfahrens: Durch den Box Score erfolgt eine Grundlage bei der Generierung neuer Hypothe-

sen für weitere Forschungsarbeiten, da beschrieben wird, was schon untersucht wurde und was noch nicht untersucht ist.

Nachteile des Verfahrens:

- Wenn viele Studien integriert werden, kann für den Leser des Reviews leicht die Übersicht verloren gehen, da eine Vielzahl von unterschiedlichen Studien mit verschiedenen potenziellen Moderatorvariablen integriert werden müssen.
- Einzelne Studien können in der Vielzahl der erfassten Studien „untergehen“, sodass relevante Studienergebnisse verschwinden.
- Analog zur Aggregation von narrativen Aussagen einzelner Probanden in einer Studie ist es schwierig, im Rahmen einer Meta-Analyse die Ergebnisse aus den einzelnen Studien zu einer Gesamtaussage zu aggregieren.
- In der Analyse werden die Stichprobengrößen der einzelnen Studien nicht berücksichtigt. Studien mit wenigen Teilnehmenden gehen genauso stark in das Gesamtergebnis ein wie Studien mit mehreren tausend Teilnehmenden.
- Unbeachtet bleiben bei der Interpretation die Größe der Mittelwertsdifferenz und daher auch die Effektgröße. Eine Studie mit einer kleinen Effektgröße wird gleich einer Studie mit einer hohen Effektgröße im Review berücksichtigt.
- Es erfolgt eine (rein qualitative) Zusammenfassung der einzelnen Studien aufgrund der p-Werte in den Kategorien signifikant ja / nein. Problematisch ist, dass das Ausmaß der Signifikanz unbeachtet bleibt. Die Zusammensetzungen der Box Counts können heterogen sein, da bei gerichteten Hypothesen nicht zwischen $p = .049$ und $p < .0001$, bzw. zwischen $p = .50$ und $p = .99$ unterschieden wird.
- Zusammenhänge zwischen abhängigen Variablen und möglichen Moderatorvariablen innerhalb eines Reviews sind bei vielen Studien relativ komplex. So können diese Zusammenhänge nur als Hypothesen für weitere Untersuchungen definiert und nicht statistisch überprüft werden.

Als Folge dieser Nachteile sind „reine“ narrative Reviews eher selten geworden. Heute werden Reviews oft mit Meta-Analysen vom Typ II kombiniert (quantitative Reviews).

2.3 Meta-Analyse (Typ II)

Hierbei handelt es sich um eine Aggregation von Ergebnissen aus Studien mit gleicher / ähnlicher Fragestellung zu aggregierten statistischen Kennwerten.

Historisch gesehen gab es schon früh erste Meta-Analysen, nur hießen sie noch nicht so:

- Roger Cotes berechnete im 18. Jahrhundert gewichtete Mittelwerte der Abweichungen verschiedener Astronomen, um die exakten astronomischen Daten besser schätzen zu können.
- Sir Karl Pearson (Entwickler der Produkt-Moment-Korrelation) kombinierte die Ergebnisse verschiedener Korrelationsstudien, indem er die Korrelationskoeffizienten nach einer Transformation nach Fisher (Fishers Z) mittelte.
- Hans Jürgen Eysenck publizierte 1952 seine Psychotherapieerfolgsstudie. Er ging davon aus, dass Psychotherapie bei Patienten keine größere Wirkung zeige als Spontanremission.
- Der Begriff Meta-Analyse wurde bei den ersten systematischen Zusammenfassungen von Therapieeffektstudien verwendet (Smith & Glass, 1977).

Vorteil dieses Verfahrens ist eine bessere Schätzung der „wahren“ Effekte durch quantitative Auswertung / Zusammenfassung mehrerer Studien. Durch die erhöhte Gesamtstichprobe sinkt die Gefahr von Fehlern erster und zweiter Art. Außerdem können „Ausreißerstudien“ entdeckt werden, die dem allgemeinen Erkenntnisstand (mittleren Effekt) widersprechen. Dies kann weitere Forschung initiieren. Auch kann durch die Berechnung von Konfidenzintervallen über Studien hinweg, z. B. im Rahmen einer Therapieerfolgsstudie, ermittelt werden, was der „normale“ und somit zu erwartende Therapieeffekt ist. Durch Untergruppenanalysen (Subgruppenanalysen) kann auf Basis einer ausreichend großen Subgruppenstichprobe untersucht werden, ob bestimmte Bedingungen unterschiedlich erfolgreich sind. Ergeben sich möglicherweise durch Therapeutinnen größerer Therapieerfolge als durch Therapeuten? Unterscheiden sich die verschiedenen Therapieschulen (Verhaltenstherapie, systemische Therapie, Tiefenanalyse etc.) in diversen Erfolgsparametern?

Bei einer Meta-Analyse werden die folgenden Arbeitsschritte durchgeführt:

- Identifizierung und Präzisierung der Fragestellung sowie ausführliche Literaturrecherche,
- Erfassung und Kodierung der Studien,

- Ermittlung der Effektgrößen,
- Analysieren der Meta-Analyse-Daten,
- Interpretation und Präsentation der Ergebnisse.

Diese Punkte werden im Folgenden ausführlicher besprochen.

Identifizierung und Präzisierung der Fragestellung sowie ausführliche Literaturrecherche: Bevor mit einer Literaturrecherche begonnen wird, muss die Fragestellung explizit dargestellt werden. Sie kann differenziert (z. B. Wie ist der Erfolg einer stationären Verhaltenstherapie bei jüngeren Borderline-Patientinnen?) oder breit gefasst sein (z. B. Wirkt Psychotherapie?). Allerdings kann eine zu breite Fragestellung zu einer „Flut“ passender Veröffentlichungen führen, sodass nach einer ersten Literaturrecherche die Fragestellung möglicherweise nochmals präzisiert werden muss. Die Auswahl der geeigneten Suchbegriffe ist ebenfalls bedeutend für die Suche nach relevanten Studien. Nach der ersten Sichtung der gefundenen Literatur muss die Suche evtl. mit neuen Schlagworten durchgeführt werden. Ziel der Literaturrecherche ist es, den aktuellen Wissensstand zu erfassen.

Die Suche wird immer durch das Problem der sog. Publikationsverzerrungen (File-Drawer-Problem) erschwert. Darunter wird die Tendenz bei Autoren und Herausgebern verstanden, bevorzugt erwartungskonforme Studien mit signifikanten Ergebnissen zu veröffentlichen. Studien mit statistisch nicht bedeutsamen Ergebnissen oder mit Daten, die den allgemeinen Grundannahmen widersprechen, verschwinden hierbei oft „in der Schublade“.

Diplomarbeiten, Promotionsarbeiten und Habilitationsschriften sind, bzw. waren früher, oft nur schwer erreichbar. Diese Publikationen werden als „graue Literatur“ bezeichnet. Forschungsberichte oder institut-sinterne Papiere sind oft der Öffentlichkeit nicht zugänglich oder können nur bei den jeweiligen Instituten direkt angefordert werden (s. z. B. für das Institut für Psychologie der Universität Freiburg unter <http://portal.uni-freiburg.de/psychologie/forschung/fobe90.html>). Diese Arbeiten und Berichte werden wie auch manche Zeitschriften nicht immer in den medizinischen oder psychologischen Datenbanksystem (Medline oder Psyndex) aufgelistet; es sollte daher so viel der „grauen Literatur“ wie möglich aufgenommen werden. Wie kann sie nun dennoch erfasst werden? Hierzu gibt es u. a. folgende Möglichkeiten:

- Die bekannten Forscher / Forschungsgruppen zur untersuchten Fragestellung anschreiben und um evtl. vorhandene unveröffentlichte Information bitten.
- Auch außerhalb der „festgetretenen Pfade“ (Datenbanken) suchen, z. B. im Internet, auf Tagungen, jeweilige Fachverbände anschreiben etc.
- In Universitätsbibliotheken nach unveröffentlichten Arbeiten suchen. Dies ist heute durch Internetzugang und Suchmaschinen (z. B. www.google.de) möglich.
- Spezielle Kliniken (z. B. für Asthma-Patienten) oder Zielgruppen (z. B. Selbsthilfegruppe) können angeschrieben werden, wobei hier nicht nur Manuskripte, sondern auch Stations- oder Klinikstatistiken Datengrundlage für Meta-Analysen sein können.

Kodierung der Studien: Bei der Erfassung der einzelnen Studien ist ein umfangreiches Kodierungsschema notwendig, damit diese korrekt erfasst und verglichen werden können. Dieses Kodierungsschema bewertet die Studien nach einer Vielzahl von Kriterien. Es muss vor der Datenerhebung entwickelt werden und sollte ausführlich getestet sein. Ratsam sind ein Training der Beurteiler (Rater) und die Bewertung jeder Studie durch mindestens zwei unabhängige Rater. Je differenzierter das Bewertungsschema, desto besser ist die Meta-Analyse auszuwerten.

Grawe, Donati und Bernauer (1994) haben in ihrer großen Therapieerfolgsstudie Güteprofile wissenschaftlicher Veröffentlichung zur Wirkung von Psychotherapie erstellt. Die Studien wurden anhand von acht *Gütekriterien* beurteilt:

1. Klinische Relevanz
2. Interne Validität
3. Güte und Information
4. Vorsicht bei der Interpretation
5. Reichhaltigkeit der Messung
6. Güte der Auswertung
7. Reichhaltigkeit der Ergebnisse
8. Indikationsrelevanz

Anhand der Beurteilungen wurden *Güteprofile* erstellt. Sie wurden grafisch aufbereitet, sodass ein direkter und schneller Vergleich der Studien bzgl. der Gütekriterien möglich ist.

Ermittlung der Effektgrößen: Über Effektgrößenberechnungen werden die in einzelnen Studien gefundenen Effekte standardisiert, sodass ein Vergleich der Ergebnisse möglich ist. Für die Vielzahl von Effektgrößenmaßen (η^2 , R^2 , r , d ...) werden beispielhaft zwei Effektgrößen dargestellt:

- standardisierte Mittelwertsdifferenzen (d-Maß), für intervallskalierte Merkmale
- Odds Ratio (OR) für dichotome Merkmale.

Das **d-Maß nach Cohen** (1988) ist die an der Streuung standardisierte Differenz zweier Mittelwerte:

$$d = M_1 - M_2 / SD$$

Hiermit lassen sich zwei Gruppenmittelwerte vergleichen, sodass z. B. Therapieerfolge in verschiedenen Studien eingeschätzt werden können. Zur *Bewertung* des d-Maßes schlägt Cohen Richtwerte vor:

- $d = 0,2$ (kleiner Effekt),
- $d = 0,5$ (mittlere Effekt),
- $d = 0,8$ (großer Effekt).

Sie sollten nur verwendet werden, wenn keine Effektgrößen aus vergleichbaren Studien vorliegen. Bei der Ermittlung des d-Maßes für abhängige Stichproben stehen verschiedene *Streuungen* (SD) zur Auswahl:

- SD der Differenzwerte nach Cohen (1988) zur Ermittlung der „standardized response mean“ (SRM),
- SD der Prä-Werte nach Kazis, Anderson und Meenan (1989) zur Ermittlung der „standardized effect size“ (SES),
- gepoolte SD der Prä-Post-Werte (Hartmann et al. 1992),
- SD der Prä-Werte aller Gruppen (Grawe et al. 1994).

Da es verschiedene Berechnungsmöglichkeiten gibt, sollte in einer Publikation die Berechnung immer transparent dargestellt werden. Mehr zur Auswahl der relevanten Streuung bei Leonhart (2004b).

Das **Odds Ratio** beschreibt das Verhältnis zweier Chancen (Odds) zueinander. Hierbei wird z. B. verglichen, wie groß die Chance ist, eine Klausur zu bestehen, wenn eine Lerngruppe besucht wurde und wenn nicht. Das Verhältnis dieser beiden Chancen gibt den Vorteil an, der durch den Besuch der Lerngruppe gegeben ist. Somit bedeutet eine Odds Ratio von 1 „kein Unterschied“ zwischen den beiden Bedingungen. Ein Wert größer als 1 beschreibt ein „Risiko“ für ein Ereignis. Ein Wert kleiner als 1 beschreibt eine „Schutzfunktion der Gruppenzugehörigkeit“ vor einem Ereignis.

Diese Berechnung soll am folgenden *Beispiel* deutlich werden (Tabelle 1). Hierbei wird das Ergebnis eines Treatments im Verhältnis zu einer Kontrollgruppe interpretiert.

Tabelle 1: Beispieldatensatz für die Berechnung von Odds Ratio

	Misserfolg / Tod	Erfolg / Überlebende
Treatmentgruppe	28 a	176 b
Kontrollgruppe	51 c	151 d

Odds Ratio (OR) wird nach folgender Formel berechnet:

$$OR = a*d/b*c = 28*151/176*51 = 0.47$$

Die Wahrscheinlichkeit, als Mitglied der Treatmentgruppe zu sterben, ist ungefähr halb so groß wie in der Kontrollgruppe. Das Treatment hat somit eine Schutzfunktion.

Analysieren der Meta-Analyse-Daten: Generell werden die Daten von Meta-Analysen analog zur „Standardstatistik“ ausgewertet, wobei die folgenden Punkte berücksichtigt werden sollten:

- unterschiedliche Stichprobengrößen,
- stark differierende Effektgrößen (Homogenitätstests für die Effektgrößen, ggf. Analysen in Subgruppen),
- hierarchische Strukturen in den Daten (d.h. Daten sind in sich abhängig),
- Abhängigkeiten der einzelnen Effektgrößen innerhalb der Studien,
- Publikationsverzerrung (publication bias).

Interpretation und Präsentation der Ergebnisse: Es besteht die Gefahr, dass aufgrund der breiten Datenbasis aus vielen Studien die Ergebnisse als „unumstößliche Wahrheit“ interpretiert werden, da große Datengrundlage „Sicherheit“ bei der Interpretation vortäuscht. Auch Zusammenhänge bei einer Meta-Analyse sind nur Korrelationen und müssen dementsprechend interpretiert werden (Vorsicht bei Aussagen über Kausalität!).

Folgende *methodische Probleme* können bei der Durchführung von Meta-Analysen auftreten und die Validität der Ergebnisse gefährden:

- Durch die Verwendung unreliabler Messinstrumente können die ermittelten Effektgrößen den wahren Effekt unterschätzen. Sind die Reliabilitätskennwerte bekannt, ist eine Korrektur möglich.
- Durch eingeschränkte Varianzen (restriction of range) werden die wahren Effektgrößen ebenfalls unterschätzt. Auch hier ist eine Korrektur möglich, wenn die Populationsvarianzen bekannt sind.
- Fehlende Angaben (z. B. kein exakter p-Wert) in den Publikationen erschweren die Auswertung. Gegebenenfalls kann bei den Autoren der Studie nachgefragt werden, was allerdings einen hohen Arbeitsaufwand ergibt.
- Gefährdung der Interraterobjektivität: Wenn bei einer Meta-Analyse die Studien durch verschiedene Reviewer nach nicht identischen Kriterien beurteilt werden, sind die Bewertungen nicht vergleichbar. Auch können sich die Urteile der Rater während Untersuchung verändern (rater drift). Ein Rater könnte z. B. im Lauf der Zeit die Kriterien für den Erfolg einer Therapiemaßnahme lockern. Da die Hypothesen und Ziele der Meta-Analyse transparent sind, werden die Urteile der Rater unbewusst bei der Bewertung in die „richtige Richtung“ beeinflusst. Dies ist durch klare Aufzeichnung der Kodierung, Training der Reviewer, Kodierung über viele Items und Bewertung durch mehrere Reviewer vermeidbar.
- Bei Meta-Analysen über viele Studien und viele potenzielle Prädiktoren besteht durch die große Anzahl von statistischen Prüfverfahren eine erhöhte Gefahr der Alpha-Fehler-Inflationierung (capitalization of chance). Dieses Problem kann durch Bonferroni-Korrektur, multivariate Analysen und zielgerichtete Testung weniger Hypothesen umgangen werden.
- Ergebnisse von Meta-Analysen können verzerrt sein (Bias), falls die ermittelten Effektgrößen der verfügbaren Studien nur teilweise in die Meta-Analyse einfließen (z. B. aus inhaltlichen oder methodischen Gründen, mangelnder Relevanz für die Kernfragestellung der Meta-Analyse). Unbeabsichtigt können eher die „gewünschten“ Ergebnisse in die Studie eingehen, während die „unerwünschten“ Studien eher als irrelevant betrachtet und ausgeschlossen werden.
- Die schon erwähnte Publikationsverzerrung (File-Drawer-Problem) kann verringert werden, indem versucht wird, unpublizierte Studien mit in die Meta-Analyse einzubeziehen (graue Literatur, z. B. Diplomarbeiten, Promotionsarbeiten). Durch einen Vergleich der Effektgrößen von publizierten und nichtpublizierten Studien kann dieser Bias ermittelt und bei der Ergebnisdarstellung / -interpretation berücksichtigt werden.
- Fehlende Werte sind auch bei Meta-Analysen ein Problem, wobei verschiedene Methoden zum Umgang hiermit in der aktuellen Literatur diskutiert werden. Gerade Verzerrungen, die auf den Schwund von Probanden in den einzelnen Primärstudien zurückgehen, potenzieren sich hierbei.
- Oft ist die statistische Unabhängigkeit der Ergebnisse nicht gegeben, weil:
 - verschiedene Effektgrößen in einer Gruppe mit ähnlichen Variablen (z. B. Fragebögen) ermittelt werden,
 - verschiedene Effektgrößen durch den Vergleich von mehreren Interventionsgruppen mit nur einer einzigen Kontrollgruppe ermittelt werden,
 - verschiedene Effektgrößen innerhalb einer Studie für unterschiedliche (Teil-)Stichproben ermittelt werden,
- verschiedene Studien zu einer Thematik von einer Forschergruppe erhoben werden.

Um reliable Effektgrößen zu erhalten, muss einerseits die Stichprobengröße berücksichtigt und andererseits bei der Auswertung geprüft werden, ob die Streuungen der erhobenen Merkmale zwischen den einzelnen Studien homogen sind (Varianzhomogenität). Bei Meta-Analysen mit wenigen Studien und geringen Stichprobenumfängen besteht die Gefahr, dass wegen zu geringer Teststärke die Varianzheterogenität nicht entdeckt wird. Diese muss bei der Berechnung der Meta-Analyse und der Darstellung der Studienergebnisse aber berücksichtigt werden.

- Ein weiteres Problem ist, dass Studien mit randomisierter Zuweisung (Experimente) und Studien ohne randomisierte Zuweisung (Quasi-Experimente) in die Meta-Analyse eingehen können. Sind diese Studien wirklich vergleichbar? Liegen ausreichend viele Studien vor, sollten zwei Meta-Analysen (randomisierte Studien vs. nichtrandomisierte Studien) durchgeführt werden.

Generell gilt für die methodische Qualität einer Meta-Analyse, dass sie von der Qualität der Primärstudien abhängt. Ist die Qualität der Daten in den Primärprojekten schlecht, kann sie nicht durch eine Meta-Analyse verbessert werden („garbage-in, garbage-out“).

Ferner muss hinterfragt werden, ob in den einzelnen Studien die jeweiligen Zielkonstrukte vollständig erhoben worden sind (Konstruktvalidität, z. B. Therapieerfolg oder nur die Qualität der Patient-Therapeut-Interaktion). Falls Studienmerkmale unreliabel oder nur mit einer von mehreren möglichen Untersuchungsmethoden erhoben werden, ist dies problematisch. Auch hängen die untersuchten Variablen oft mit anderen Konstrukten zusammen (z. B. ist eine Rehabilitationsmaßnahme in Deutschland meistens eine stationäre Maßnahme).

Wenn die Heterogenität der eingeschlossenen Populationen, Treatments, Erhebungsinstrumente, Settings und Messzeitpunkte sehr groß ist, besteht der Verdacht, dass unterschiedliche Konstrukte miteinander verglichen werden (Äpfel-und-Birnen-Problem, z. B. durch zu unterschiedliche Operationalisierungen). Eine Lösung besteht darin, die Einschlusskriterien für die einzelnen Studien zu erhöhen. Zu große Heterogenität bei ermittelten Effektgrößen der Primärstudien ist ebenfalls ein Grund, an den Ergebnissen der Meta-Analyse zu zweifeln.

2.4 Meta-Analysen des Typs III und IV

Abschließend noch die Frage nach der Durchführbarkeit von Meta-Analysen vom Typ III und IV. Da diese Form der Meta-Analyse die Bereitstellung und Überlassung der Rohdaten durch die jeweiligen Projektleiter der ursprünglichen Projekte (Primärprojekte) voraussetzt, ist die Datensammlung schwierig. Eigene Erfahrungen aus einem Projekt haben gezeigt, dass die Hürden eher vielfältig sind und meistens nur schwer überwunden werden können (mehr hierzu s. Leonhart/Maurischat 2004). Wegen der Ermittlung von studienübergreifenden quantitativen Effekten und Effizienzen gelten Meta-Analysen gegenüber qualitativen Reviews als höherwertig und werden im Moment verstärkt angewendet (Cochrane/Silagy 1999).

Zusammenfassung: Über Reviews und Meta-Analysen werden Ergebnisse aus einzelnen Studien zusammengefasst und kompakt dargestellt. Dieses Vorgehen ist wegen einer exponentiell anwachsenden Anzahl von publizierten Studien notwendig geworden, da ein einzelner Wissenschaftler sonst den Überblick über den aktuellen Forschungsstand leicht verlieren kann. Während Reviews die Studienergebnisse eher deskriptiv zusammenfassen, werden bei Meta-Analysen über mehrere Studien gemittelte Effektgrößen berechnet. Trotz aller Probleme bei der Durchführung von Meta-Analysen und der Interpretation der Ergebnisse werden zukünftig Meta-Analysen verstärkt eingesetzt.

Übungsaufgaben

1. Was unterscheidet eine Review von einer Meta-Analyse?
2. Warum steigt die Notwendigkeit der Durchführung einer Meta-Analyse?
3. Welches Vorgehen wird mit den Begriffen Box Score oder Vote Counting beschrieben?
4. Definieren Sie das Effektgrößenmaß Odds Ratio.
5. Definieren Sie das Effektgrößenmaß d von Cohen.
6. Warum ist die Durchführung eines Reviews ohne statistische Auswertung gerade bei einer Vielzahl von Studien nicht sinnvoll?
7. Welche Probleme entstehen bei der Literaturrecherche für eine Meta-Analyse bezüglich der „grauen Literatur“?

Antworten:

1. Bei einem Review findet nur eine deskriptive Beschreibung der Studienergebnisse statt, während bei einer Meta-Analyse die statistischen Kennwerte der einzelnen Studien in ein Gesamtergebnis integriert werden (z.B. Ermittlung der mittleren Effektgröße über alle Studien hinweg).
2. Innerhalb der Wissenschaft steigt die Anzahl der Publikationen und Zeitschriften exponentiell an, sodass die Flut der Informationen für einen einzelnen Wissenschaftler nicht mehr überblickbar sind.
3. Innerhalb eines Reviews werden die vorgefundenen Studien in zwei oder drei Kategorien (+/0/-) zugeordnet, je nachdem, zu welchem Ergebnis sie kommen. Anschließend wird die Anzahl der Studien pro Kategorie durch einfaches Abzählen bestimmt, wobei dann die Kategorie mit dem meisten Studien „gewonnen“ hat.
4. Das Odds Ratio beschreibt das Verhältnis zweier Chancen (Odds) zueinander. Hierbei gilt für ein Vierfelder-Schema: $OR = a*d/b*c$.
5. Das d -Maß nach Cohen ist die an der Streuung standardisierte Differenz zweier Mittelwerte:
 $d = M_1 - M_2 / SD$.
6. Der Leser des Reviews kann den Überblick über die Vielzahl integrierter Studien verlieren, sodass er nach dem Lesen des Reviews „so klug ist wie zuvor“.
7. Die graue Literatur ist nicht in den Literaturdatenbanken gelistet und meist nur schwer zugänglich.