2 Welcher Studientyp ist geeignet?

Heute hat sich in der Apotheke die Pharmareferentin einer Firma angekündigt, die sich auf Nahrungsergänzungsmittel spezialisiert hat. Die Pharmareferentin legt eine dicke Fachbroschüre auf den Tisch: „Hier finden Sie den Literaturreport für unser Mittel zur Stärkung der Immunabwehr." In der Broschüre finden Sie Zusammenfassungen zahlreicher Studien. Das sieht so aus, als wäre die Wirksamkeit wirklich umfangreich untersucht – oder doch nicht?

Die Werbung für Arzneimittel, Nahrungsergänzungsmittel und manchmal auch Kosmetika zitiert häufig Studien, die die Wirksamkeit belegen sollen. Dann ist es jedoch notwendig, sich die langen Literaturlisten einmal genau anzuschauen. Nicht selten verstecken sich hinter den Zitaten Laborversuche, sei es an Zelllinien, Tieren oder isoliertem Material, etwa menschlicher Haut. In vielen Fällen sind solche Untersuchungen im Verlauf der Entwicklung sinnvoll und notwendig.

Allerdings ist nicht automatisch gewährleistet, dass sich die Ergebnisse auch auf den Menschen beziehungsweise einen lebendigen Organismus übertragen lassen. Deshalb sind zusätzlich klinische Studien notwendig.

Allerdings gibt es bei solchen Studien erhebliche Unterschiede. Wenn es darum geht, einen ursächlichen Zusammenhang zwischen der Einnahme eines Arzneimittels oder Nahrungsergänzungsmittels und einem verbesserten Gesundheitszustand nachzuweisen oder sicher zu quantifizieren, sind nicht alle Studientypen gleich gut geeignet.

2.1 Wie sich Studientypen unterscheiden

Wer herausfinden will, um welchen Studientyp es sich handelt, findet die Bezeichnung oft schon im Titel oder in der Zusammenfassung, anderenfalls im Methodenteil der Publikation. Ansonsten können bei der Unterscheidung verschiedene Kriterien helfen (Abb. 2.1).

- Werden die untersuchten Behandlungen (= Interventionen), etwa ein bestimmtes Arzneimittel, explizit im Rahmen der Studie zugeteilt? Dann handelt es sich um eine interventionelle Studie, auch experimentelle oder klinische Studie genannt. Oder beobachtet die Studie Behandlungen oder Einflüsse (= Exposition), die auch ohne die Studie so vorgenommen würden oder vorhanden wären? Dann ist es eine Beobachtungsstudie (epidemiologische Studie).
- Gibt es in der Studie eine Kontrolle, wird das zu untersuchende Arzneimittel also zum Beispiel mit einem anderen Arzneimittel, einem Placebo oder keiner Behandlung verglichen? Dann handelt es sich um eine kontrollierte Studie, wenn nicht, um eine unkontrollierte Studie. Bei klinischen Studien ist dann auch oft von „einarmi-

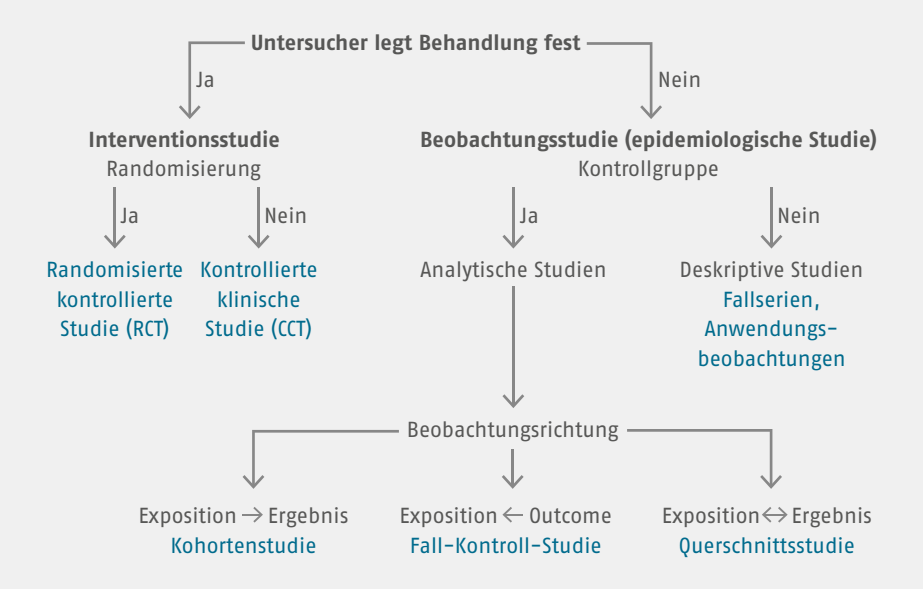

Abb. 2.1 Übersicht über die Einteilung von Studientypen in der klinischen Forschung. Nach Grimes 2002a

gen" Studien die Rede, während kontrollierte klinische Studien zwei- oder mehrarmig sein können, je nachdem, wie viele Interventionen verglichen werden. Bei den Beobachtungsstudien gibt es eine Kontrollgruppe, ebenso bei Kohortenstudien und Fall-Kontroll-Studien (sogenannte analytische Studien), die Kontrollgruppe fehlt dagegen bei Fallberichten, Fallserien oder Anwendungsbeobachtungen (deskriptive Studien).

- Falls es eine Kontrolle gibt: Werden die Patient:innen nach dem Zufallsprinzip (randomisiert) auf die Behandlungs- und Vergleichsgruppe aufgeteilt? Dann handelt es sich um eine randomisierte kontrollierte Studie, anderenfalls um eine nicht-randomisierte kontrollierte Studie.
- Erhalten die Teilnehmenden eine Behandlung und werden dann weiter beobachtet (prospektive Beobachtung), oder sucht das Forschungsteam in der Vergangenheit von Patient:innen mit bestimmten Eigenschaften nach den interessierenden Einflussfaktoren oder Behandlungen (retrospektive Beobachtung)?

Für unterschiedliche Fragestellungen nutzt die Forschung dabei typischerweise verschiedene Studientypen, die in diesem Rahmen alle ihre Berechtigung und einen Stellenwert haben (Tab. 2.1).

3 Wie verlässlich ist die Studie?

Ihre Kollegin hat einen Werbeprospekt für ein neues OTC-Präparat aufmerksam gelesen. „Das Mittel wurde in zwei randomisierten kontrollierten Studien untersucht", berichtet sie begeistert. „Dann ist es doch sicher, dass die Wirksamkeit tatsächlich erwiesen ist. Soll ich gleich das Einführungsangebot bestellen?"

Wie im letzten Kapitel ausführlich beschrieben, lässt sich nicht mit allen Arten von Studientypen im gleichen Maß belegen, dass tatsächlich ein kausaler Zusammenhang zwischen der Einnahme eines Arzneimittels und der Verbesserung des Gesundheitszustands besteht. Der Einfluss anderer Faktoren, die die Ergebnisse in klinischen Studien verzerren können, wird am besten durch randomisierte kontrollierte Studien reduziert, die deshalb auch als „Goldstandard" in der Beurteilung von Fragen zu Prävention und Therapie gelten (▸Kap. 2.3.5). Denn die randomisierte Zuteilung der Teilnehmenden auf Behandlungs- und Kontrollgruppe gewährleistet, dass die Gruppen strukturell ähnlich sind, es also faire Ausgangsbedingungen für den Vergleich gibt.

Allerdings können bei randomisierten kontrollierten Studien auch noch weitere Arten von Bias auftreten (○Abb. 3.1). Deshalb sind weitere Maßnahmen nötig. Allerdings kann ein unsachgemäßes Vorgehen dazu führen, dass diese Prinzipien nicht zum gewünschten Erfolg führen, Verzerrungen durch systematische Fehler also nicht im gewünschten Ausmaß reduzieren. Deshalb ist es wichtig, bei der Bewertung einer randomisierten kontrollierten Studie immer auch die Details der Umsetzung, sprich das Bias-Risiko der Studie und damit die interne Validität der Studie, zu prüfen. Die notwendigen Angaben dazu sollten sich im Methodenteil der Publikation finden.

Allerdings sind die Beschreibungen von Studien nicht immer detailliert genug, um alle Einzelheiten zu bewerten. Daraus lassen sich keine Schlussfolgerungen für die interne Validität einer Studie ziehen, sondern nur über die Berichtsqualität. Manchmal können außer der Publikation in einer wissenschaftlichen Fachzeitschrift auch noch weitere Quellen wie Studienregister helfen (▸Kap. 1.3).

In manchen Fällen ist die Aussagekraft einer Studie auch nicht dadurch eingeschränkt, dass die Ergebnisse verzerrt sind, sondern dass sie sich nicht auf andere Umstände übertragen lassen. Das ist etwa dann der Fall, wenn nur eine sehr eingeschränkte Gruppe von Patient:innen an der Studie teilgenommen hat und keine patientenrelevanten Endpunkte erhoben wurden. Das ist kein Problem der internen, sondern der externen Validität der Studie (▸Kap. 5).

Bei der Prüfung der internen Validität ist auch wichtig zu prüfen, wie stark das Fehlen oder die unzureichende Umsetzung einer Maßnahme tatsächlich das Bias-Risiko erhöht. Bei fehlender Verblindung können die Auswirkungen auf die Erhebung je nach Endpunkt sehr unterschiedlich sein (▸Kap. 3.3).

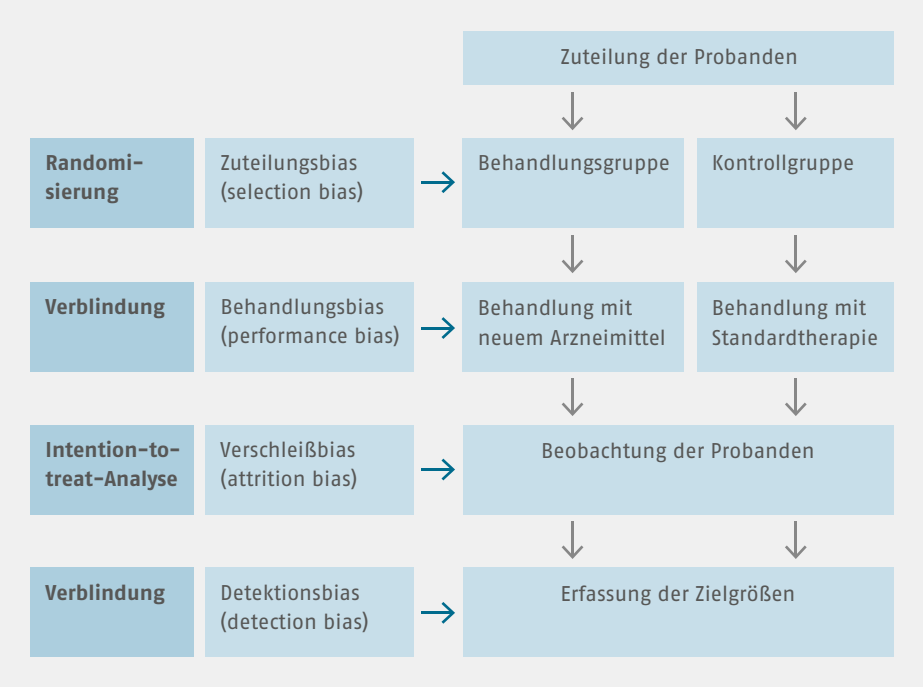

Abb. 3.1 Mögliche Quellen für Bias in randomisierten kontrollierten Studien und Möglichkeiten der Abhilfe. Nach Greenhalgh 1997

3.1 Wie Vergleiche aussehen können

Mit einer Kontrollgruppe lässt sich in einer Studie ermitteln, welcher Teil des Behandlungserfolges auf das zu untersuchende Medikament und welcher Teil auf andere Einflüsse zurückzuführen ist. Fehlt eine Kontrollgruppe, ist das nicht immer klar.

Bei Erkrankungen mit einer hohen Tendenz zur Selbstheilung, etwa einer Erkältung, kann sich das Krankheitsbild auch von selbst verbessert haben. Bei vielen chronischen Erkrankungen können Verlauf und Schweregrad individuell stark variieren. Besonders bei Erkrankungen, die in Schüben verlaufen, etwa bei einer rheumatoiden Arthritis oder bei chronisch-entzündlichen Darmerkrankungen, folgen auf Phasen mit einer hohen Krankheitsaktivität in der Regel auch ohne äußere Einflüsse Zeiträume, in denen die Patient:innen nahezu beschwerdefrei ist (Remissionsphasen).

Schließlich gibt es außer der medikamentösen Behandlung in vielen Fällen auch weitere Faktoren, die den Krankheitszustand oder die Symptome beeinflussen: So können sich bestimmte Erkrankungen und Zustände etwa durch psychische Einflüsse, zum Beispiel ärztliche oder pflegerische Zuwendung, verbessern. Auch Ernährungs- und Bewegungsverhalten, ein Rauch-Stopp oder andere unterstützende Maßnahmen wie Bettruhe können Krankheitsverläufe je nach Erkrankung ebenfalls günstig beeinflussen.

Wichtig ist es deshalb, dass sich Behandlungs- und Kontrollgruppe tatsächlich nur in der zu untersuchenden Intervention, also zum Beispiel dem zu untersuchenden Arzneimittel, unterscheiden. Damit sich das überprüfen lässt, muss die Behandlung in der Publikation der Studie ausreichend detailliert beschrieben sein.

Zugang zu den unverblindeten Studiendaten hat. Das DSMB sollte die Daten nach vorher festgelegten statistischen (Stopp-)Regeln bewerten, die die beschriebenen Probleme berücksichtigen.

In der Studienpublikation lohnt es sich deshalb, darauf zu achten, ob das Vorgehen im Methodenteil entsprechend beschrieben ist. Aufmerksamkeit ist besonders dann geboten, wenn insgesamt weniger Patient:innen als geplant an der Studie teilgenommen haben, das aber in der Publikation nicht weiter diskutiert wird.

4.5 Wie sich die Ergebnisse von Nichtunterlegenheitsstudien bewerten lassen

Die Studien, um die es bisher ging, beschäftigten sich mit der Frage, ob ein neues Arzneimittel besser wirkt als die bisherige Standardtherapie. Das bezeichnet man auch als Überlegenheitsstudie. Für die statistischen Tests (▸ Kap. 4.3.1) in einer solchen Studie lautet die Nullhypothese (bei zweiseitiger Fragestellung): Die beiden Mittel wirken gleich gut, es gibt keinen Unterschied. Die Alternativhypothese heißt: Es gibt einen Unterschied zwischen den beiden Mitteln.

Nun geht es manchmal aber nicht darum, die Überlegenheit eines neuen Mittels nachzuweisen, sondern zu zeigen, dass es gleich gut wie die Standardtherapie oder ihr zumindest nicht unterlegen ist. Das kann etwa dann der Fall sein, wenn zu vermuten ist, dass das neue Mittel ungefähr gleich gut wirkt wie der ältere Wirkstoff, aber weniger Nebenwirkungen hat, für Patient:innen mit Kontraindikationen gegen den älteren Wirkstoff geeignet ist, leichter anzuwenden oder deutlich preisgünstiger ist. Für viele Indikationen gibt es außerdem inzwischen eine gut wirksame Standardtherapie, sodass es sehr schwer wäre, eine Überlegenheit nachzuweisen.

In solchen Fällen wird gelegentlich ein Studiendesign gewählt, das je nach konkretem Design als Äquivalenzstudie oder Nichtunterlegenheitsstudie bezeichnet wird. Genauer formuliert bedeutet Äquivalenz, dass das neue Arzneimittel weder besser noch schlechter wirkt als die Standardtherapie, während Nichtunterlegenheit heißt: Zumindest nicht schlechter, der Fall „besser wirksam“ ist aber auch möglich. (Bio-)Äquivalenzstudien kommen beispielsweise bei der Zulassung von Generika zum Einsatz.

Würde man bei Nichtunterlegenheitsstudien einfach die gleichen Kriterien anwenden wie bei Überlegenheitsstudien, könnte es allerdings leicht zu Fehlschlüssen kommen. Deshalb gibt es bei Nichtunterlegenheitsstudien spezielle Anforderungen an Planung und Auswertung.

Das beginnt bereits bei der Formulierung der Hypothesen, die bei Nichtunterlegenheitsstudien umgekehrt wie bei einer Überlegenheitsstudie aufgestellt werden. Die Nullhypothese heißt also: „Das alte Arzneimittel wirkt besser als das neue“, die Alternativhypothese: „Das neue Arzneimittel wirkt nicht schlechter als das alte.“

Beispiel Eine Studie soll nachweisen, dass ein neues Mittel genauso gut wirkt wie die bisherige Standardtherapie. In einem Überlegenheitsdesign wäre es leicht, zu diesem Ergebnis zu kommen, nämlich dann, wenn die Ergebnisse zwischen den Teilnehmenden so stark streuen, dass ein statistischer Test keinen signifikanten Unterschied zwischen den Gruppen findet. Im Grunde wäre die Aussage dann aber nur: „Die Überlegenheit des alten Arzneimittels lässt sich nicht nachweisen.“ Ob das neue Mittel tatsächlich gleichwertig beziehungsweise nicht unterlegen ist, lässt sich so aber nicht sicher nachweisen.

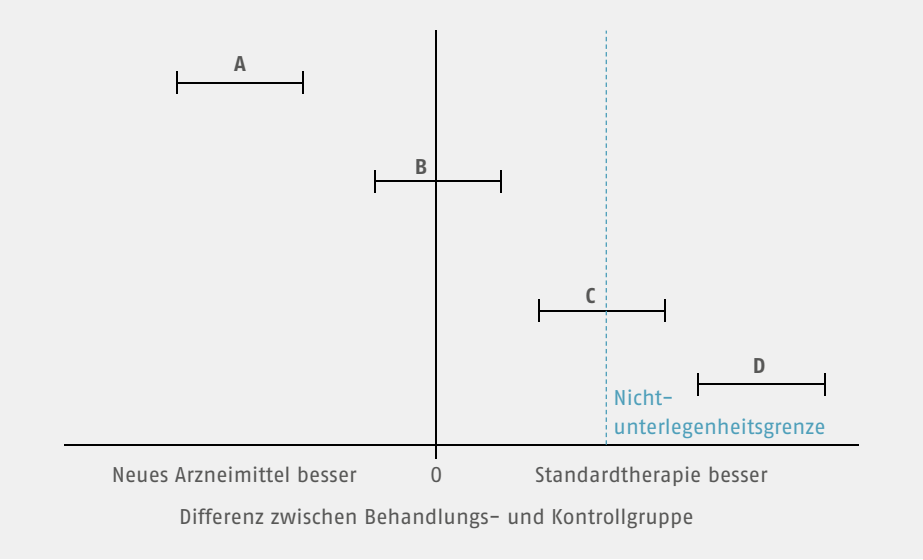

Abb. 4.3 Beispiel für die Beurteilung von Nichtunterlegenheit. Gezeigt sind jeweils die Konfidenzintervalle für den Unterschied der Therapieeffekte zwischen Behandlungs- und Kontrollgruppe. In den Fällen A und B ist Nichtunterlegenheit des Mittels in der Behandlungsgruppe gezeigt, da sich die Konfidenzintervalle vollständig links von der Nichtunterlegenheitsgrenze (gestrichelte Linie) befinden (zugunsten der Behandlungsgruppe). In Fall C ist keine eindeutige Beurteilung möglich, da das Konfidenzintervall die Nichtunterlegenheitsgrenze einschließt. In Fall D lässt sich nicht auf Nichtunterlegenheit des Mittels in der Behandlungsgruppe schließen, da sich das Konfidenzintervall vollständig diesseits der Nichtunterlegenheitsgrenze (zugunsten der Kontrollgruppe) befindet. Nach Piaggio et al. 2012

Für den Nachweis von Nichtunterlegenheit ist deshalb auch eine angemessene Fallzahlplanung wichtig. Eine zu geringe Fallzahl könnte dazu führen, dass ein in Wirklichkeit bestehender Unterschied aus statistischen Gründen in der Studie nicht sichtbar wird. Hat die Studie also zu wenig Teilnehmende, steigt das Risiko für einen Fehler 2. Art – man geht also fälschlicherweise davon aus, dass kein Unterschied besteht, obwohl in Wirklichkeit einer vorhanden ist.

Natürlich ist es auch wichtig zu definieren, wie viel Unterschied zwischen den beiden Mitteln noch akzeptabel ist. Diese Nichtunterlegenheitsgrenze wird vor Studienbeginn festgelegt. Sie entsteht aus klinischen Überlegungen, also: Wie viel schlechter dürfte das neue Mittel helfen, damit es für Patient:innen keinen wesentlichen Unterschied macht? Diese Überlegungen sollte das Forschungsteam in der Publikation nachvollziehbar darlegen.

Bei der Auswertung der Studiendaten gilt das Prinzip: Wenn der Unterschied zwischen Behandlungs- und Kontrollgruppe kleiner ist als die Nichtunterlegenheitsgrenze, gilt das neue Medikament als „nicht unterlegen", für Patient:innen sollte es also keinen wesentlichen Unterschied bei der Wirksamkeit machen, welches der beiden Mittel sie erhalten. Diese Auswertung erfolgt in der Regel auf der Basis der entsprechenden Konfidenzintervalle. Das CONSORT-Statement (Schulz et al. 2010) empfiehlt, die Äquivalenz- bzw. Nichtunterlegenheitsgrenzen und die Konfidenzintervalle in einer Grafik darzustellen, damit die Leser:innen die Ergebnisse leicht nachvollziehen können (Abb. 4.3).

mehr als fünf Studien? Subgruppeneffekte sind umso glaubwürdiger, je größer die Anzahl der Studien ist, die Daten für die Vergleiche liefern.

- Hat das Review-Team ein Random-Effects-Modell benutzt? Das Random-Effects-Modell (▸Kap. 6.3.1) berücksichtigt, dass sich die Therapieeffekte zwischen den Studien unterscheiden können. Zusätzlich stärkt es die Ergebnisse des Interaktionstests auf Subgruppeneffekte (▸Kap. 4.4.2), weil mit dem Random-Effects-Modell nicht so schnell statistisch signifikante Ergebnisse zu erreichen sind.

6.3.3 Sensitivitätsanalysen

Bei Sensitivitätsanalysen kann das Review-Team prüfen, wie sich methodische Entscheidungen oder methodische Eigenschaften der Einzelstudien auf den Gesamt-Effektschätzer auswirken. So kann eine Sensitivitätsanalyse zum Beispiel Studien mit hohem Bias-Risiko ausschließen, ein anderes statistisches Modell nutzen oder die Altersgrenzen für eine Subgruppenanalyse anders ziehen. Wenn das Ergebnis der Hauptanalyse und der Sensitivitätsanalysen übereinstimmen, gilt der Gesamt-Effektschätzer der Hauptanalyse als robust. Anderenfalls ist das Ergebnis eher mit Vorsicht zu bewerten.

6.4 Publikationsbias

Untersuchungen in den letzten Jahren haben gezeigt, dass die Ergebnisse von durchgeführten Studien in vielen Fällen nicht oder nicht vollständig veröffentlicht werden. Dieses Phänomen bezeichnet man auch als Publikationsbias.

Die Nichtveröffentlichung kann dabei verschiedene Dimensionen annehmen: So werden einige Studien (besonders solche mit negativen Ergebnissen) gar nicht veröffentlicht, auch nachträgliche Veränderungen bei der Wahl der Endpunkte sowie bei Subgruppenanalysen sind keine Seltenheit. Bei anderen Publikationen von Studien dagegen fehlen Daten zu einzelnen Endpunkten oder Nebenwirkungen. Das kann gravierende Auswirkungen haben.

Praxisbeispiel

Veröffentlichte Studien hatten dem Antidepressivum Reboxetin ein positives Nutzen-Risiko-Verhältnis bescheinigt. Bei der Recherche stieß ein Review-Team auf Hinweise zu unveröffentlichten Studien. Als der Hersteller die Daten auf mehrfache Nachfrage hin zur Verfügung stellte, wurde klar, dass bis zu diesem Zeitpunkt nur Studien mit positiven Ergebnissen veröffentlicht worden waren. Nach Einschluss aller Daten in die Metaanalyse ließ sich kein Nutzen mehr für Reboxetin nachweisen. Das Review-Team stellte auch fest, dass durch den Publikationsbias die Risiken des Wirkstoffs deutlich unterschätzt wurden (Eyding et al. 2010). Ähnliche Auswirkungen hatte ein Publikationsbias auch auf die Bewertung von Neuraminidasehemmern zur Therapie der Influenza (Jefferson et al. 2014).

Aus diesem Grund gibt es seit geraumer Zeit verschärfte Regeln: So müssen die meisten klinischen Studien zu Arzneimitteln vor Beginn in entsprechenden Studienregistern registriert werden. Die Registrierung der entsprechenden Studien haben auch einige medizinische Fachzeitschriften (etwa British Medical Journal, The Lancet, Journal of the American Medical Association) bereits als Voraussetzung für die Publikation der Ergeb-

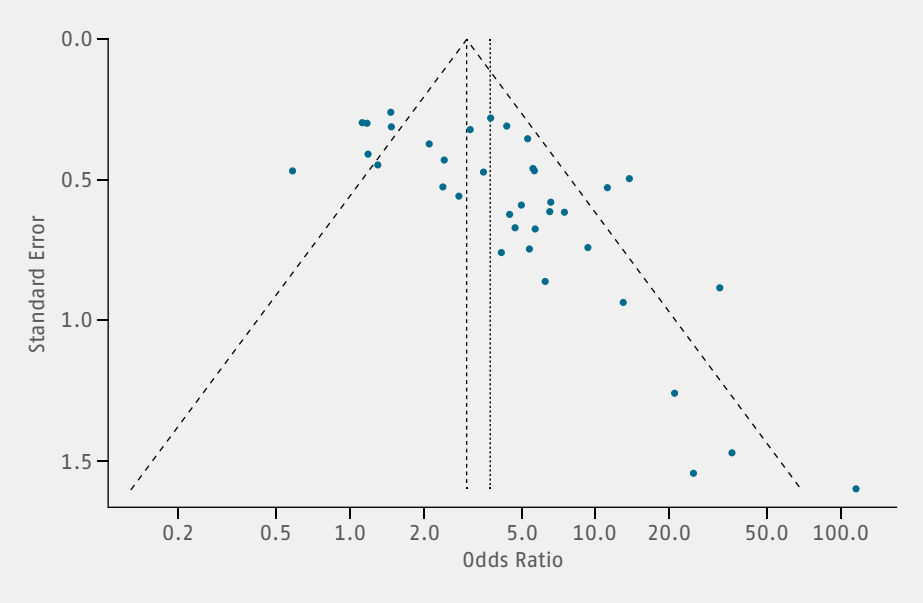

Abb. 6.7 Beispiel für einen Funnel-Plot. Aufgetragen sind die Standardfehler (Maß für die Varianz) gegen die Effektgröße (logarithmiertes Odds Ratio) der einzelnen Studien. Die Asymmetrie deutet darauf hin, dass ein Publikationsbias vorliegen könnte. Nach Schwarzer et al. 2015

nisse etabliert. Auch die Veröffentlichung der Studienergebnisse ist gesetzlich geregelt. Allerdings zeigen Untersuchungen, dass es dabei immer noch Lücken gibt oder viel Zeit bis zur Publikation vergehen kann. Verstöße gegen die Regelungen sanktionieren die zuständigen Behörden bisher allerdings nicht in allen Fällen, nicht regelmäßig und nicht zeitnah.

Für systematische Übersichtsarbeiten bedeutet das: Das Review-Team sollte sehr sorgfältig nach publizierten Studienergebnissen und in Studienregister nach durchgeführten Studien suchen (▸Kap. 6.2.2). Außerdem können auch grafische und statistische Methoden Anhaltspunkte liefern, ob möglicherweise ein Publikationsbias vorliegt.

Eine grafische Methode ist der sogenannte Funnel-Plot (englisch: funnel = Trichter). In einem Diagramm wird der Standardfehler als Maß für die Varianz der Einzelstudien gegen die jeweils beobachtete Effektgröße aufgetragen. In der Regel weisen große Studien eine geringe Varianz der Effektschätzer auf, während sich bei kleineren Studien eine größere Streuung findet. Dadurch ergibt sich im Normalfall die Darstellung eines umgekehrten Trichters: Die Spitze bilden wenige Studien mit einer großen Anzahl an Teilnehmenden und geringer Varianz, die Basis meist mehrere kleine Studien mit größerer Varianz (Abb. 6.7). Fehlen dagegen beispielsweise kleine Studien mit negativen Ergebnissen, wird der Trichter asymmetrisch.

Allerdings braucht es für aussagekräftige Ergebnisse im Funnel-Plot ausreichend viele Datenpunkte, weil sonst auch rein zufällige Effekte für die Asymmetrie verantwortlich sein können. Als Faustregel gilt dabei, dass mindestens zehn Studien vorhanden sein sollten. Gleiches gilt auch für statistische Methoden wie den Eggert-Test.

Außerdem kann es auch noch andere Gründe für einen asymmetrischen Funnel-Plot geben, die nichts mit einem Publikationsbias zu tun haben: So können kleine Studien bei