Studien richtig lesen

Teil 1: Studiendesign und Fehlerquellen

Bildnachweis: Studio Nordbahnhof / Thieme Verlagsgruppe

Niemand wird Arzt aus Leidenschaft für Literaturrecherchen. Aber um fachlich auf dem Laufenden zu bleiben, muss man mehr oder weniger regelmäßig wissenschaftliche Veröffentlichungen lesen. Hier erfahren Sie, worauf es sich zu achten lohnt – damit Sie relevante Studien möglichst schnell erkennen und die Ergebnisse beurteilen können.

Was wollen Sie von der Studie wissen? Neuer Wirkstoff entdeckt, schnellerer Test entwickelt, bessere OP-Methode etabliert – wenn man Schlagzeilen und Pressemeldungen glaubt, müsste der wissenschaftliche Fortschritt in der Medizin rasant sein. Aber schaut man genau hin, halten nicht alle Studien, was sie z. B. in der Überschrift versprechen – oder zumindest suggerieren. „Das neue Medikament senkt zwar vielleicht den Blutdruck – aber Ihr Patient will wissen, ob es sein Leben verlängert oder verbessert", sagt PD Dr. Hans-Hermann Dubben. Er forscht und lehrt am Institut für Allgemeinmedizin der Uniklinik Hamburg-Eppendorf zu evidenzbasierter Medizin.

Bekommt er eine neue Studie in die Hand, entscheidet Dubben zunächst anhand des Abstracts, ob sie ihn überhaupt interessiert und welche Frage sie ihm beantworten soll. „Will ich eine neue Therapie kennenlernen, etwas über Nebenwirkungen erfahren oder über Risikofaktoren für eine Krankheit?" Davon hängt u. a. ab, ob die Studie randomisiert oder verblindet sein muss. „Den Lebensstil, Rauchen oder Alkoholkonsum kann ich schlecht randomisieren – aber Medikamente schon eher", so Dubben. „Die Verblindung der Untersucher ist bei einem harten Endpunkt wie Gesamt-Mortalität weniger wichtig: Es gibt selten Streit, ob jemand gestorben ist oder nicht. Anders ist das bei der Einschätzung einer Depression. Ohne Verblindung kann sich dort Voreingenommenheut deutlich auf das Ergebnis auswirken." Es lohnt sich also, zunächst einen Blick auf das Studiendesign zu werfen.

Tab. 1 Studiendesigns

Unterscheidungsmerkmal	Beschreibung, Besonderheiten
zeitliche Perspektive	> retrospektiv: nachträgliche Auswertung von Daten aus Patientenakten, Krankenhausinformationssystemen u. ä. > prospektiv: Datenerfassung laufend ab Studienbeginn über einen gewissen Zeitraum, anschließend Auswertung
Beobachtungsdauer	> Querschnittsstudie: „Momentaufnahme", z. B. einmalige Messung oder Befragung > Längsschnittstudie: Betrachtung im zeitlichen Verlauf bzw. mehrfache Messung mit zeitlichem Abstand
Kontrollgruppe	> deskriptive Studie: ohne Kontrollgruppe (z. B. Fallserien, Anwendungsbeobachtungen) > analytische Studie: mit Kontrollgruppe
Blickwinkel auf Exposition und Outcome	> Kohortenstudie: Probanden werden bestimmten Gruppen zugeordnet und später das Outcome verglichen (Blick von Exposition zu Outcome), prospektiv oder retrospektiv möglich > Fall-Kontroll-Studie: Patienten mit bekanntem Outcome werden anhand vorangegangener Interventionen oder Risikofaktoren verglichen (Blick von Outcome zu Exposition), immer retrospektiv. Geeignet u. a. für seltene Ereignisse.
Interventionsgrad	> Beobachtungsstudie: kein Eingriff in die Behandlung, Exposition unabhängig von Studie vorhanden, kein Kausalitätsnachweis > epidemiologische Studie: erforscht z. B. Risikofaktoren und ihre Verteilung in der Bevölkerung, meist Beobachtungsstudie > experimentelle Studie / Interventionsstudie: Probanden werden verschiedenen Interventionen zugeordnet > kontrolliert: Studie legt Intervention fest, Kontrollgruppe vorhanden > randomisiert kontrolliert: zusätzlich zufällige Verteilung der Probanden auf die Gruppen > (doppelt) verblindet randomisiert kontrolliert: zusätzliche Verblindung der Probanden (und Behandler)
Untersuchungsobjekt	> Laborexperiment > Tierversuch > klinische Studie (am Menschen)
Phasen der Medikamenten-entwicklung	> Phase 0: Erforschung der Pharmakokinetik, -dynamik, etc. > Phase I: zusätzlich Erforschung der Verträglichkeit und Sicherheit > Phase II: Überprüfung des Therapiekonzepts (Proof of Concept, IIa) bzw. Finden der geeigneten Dosis (Dose Finding, IIb) > Phase III: Wirksamkeitsnachweis für die Marktzulassung (nach der Zulassung heißen laufende Studien Phase IIIb-Studien) > Phase IV: Studie mit bereits zugelassenem Medikament in der zugelassenen Indikation, verlangt von Zulassungsbehörden z. B. zur Feststellung sehr seltener Nebenwirkungen oder initiiert von Hersteller zu Marketingzwecken
Studienziel	> Überlegenheitsstudie: soll Überlegenheit z. B. eines Präparats gegenüber der Alternative oder Plazebo zeigen > Äquivalenzstudie: soll zeigen, dass z. B. ein neuer Wirkstoff nicht schlechter oder besser wirkt als ein anderer > Nicht-Unterlegenheitsstudie: soll zeigen, dass z. B. ein neuer Wirkstoff nicht schlechter ist als ein anderer. Die beiden letzten verwendet man, falls ein Vergleich mit Plazebo unethisch ist, oder wenn die neue Behandlung nicht besser wirkt, aber weniger Nebenwirkungen hat als die Vergleichsbehandlung.
Zusammenfassung von Einzelstudien	> Review / Übersichtsarbeit: Auswertung der relevanten aktuellen Literatur (bei „systematischem" Review nach festem Standard) > Metaanalyse: systematisches Review, das die Ergebnisse der Einzelstudien mithilfe statistischer Verfahren zusammenfasst

Was das Studiendesign erwarten lässt

Verschiedene Einteilungen möglich Um eine Veröffentlichung richtig einzuordnen, kann man zunächst im Titel oder Abstract schauen, um was für eine Art von Studie es sich handelt. Davon hängt nämlich z. B. ab, ob sich mit ihr wirklich eine Kausalität beweisen lässt – unabhängig von den vielleicht beeindruckenden Zahlen im Ergebnisteil. Einen Überblick über die häufigsten Studiendesigns gibt ▶ Tab. 1. Da diese Charakterisierungen verschiedene Aspekte betrachten, wird sich eine bestimmte Studie an mehreren Stellen der Tabelle wiederfinden. Die wichtigsten Studienarten stellen wir im Folgenden genauer vor.

Randomisiert-kontrollierte Studie

Standard für Wirkungsnachweise Die randomisiert-kontrollierte Studie (randomized controlled trial, RCT) gilt als Goldstandard für die experimentelle Überprüfung einer Fragestellung. Häufig soll eine Kausalität belegt oder widerlegt werden, z. B. der Wirkungsnachweis von Medikamenten. RCTs beruhen auf folgenden Prinzipien:

- Randomisierung: Die Probanden werden vor Versuchsbeginn nach dem Zufallsprinzip mind. 2 Gruppen zugeordnet. Durch die zufällige Zuteilung werden bekannte und unbekannte Einfluss- bzw. Störfaktoren gleichmäßig auf die Gruppen verteilt, sodass Unterschiede zwischen den Gruppen wahrscheinlich auf die Intervention zurückzuführen sind [1]. Falls bekannt ist, dass ein bestimmter Faktor (z. B. Geschlecht, Alter) das Studienergebnis deutlich beeinflusst, kann man diesen auch vor der Randomisierung gleichmäßig auf die Gruppen verteilen (stratifizierte Randomisierung) [2].
- Kontrolle: Eine Gruppe dient als Kontroll- / Plazebogruppe, d. h. sie bekommt keine Behandlung, eine Scheinbehandlung oder die Standardtherapie. Eine oder mehrere andere Gruppen sind die Interventions- / Verumgruppe(n) und erhalten die zu unter-

Tab. 2 Häufige systematische Fehlerquellen

Fehlerquelle	Erklärung
Fallzahl zu klein	Bei zu kleinen Stichproben können vorhandene Unterschiede evtl. nicht nachgewiesen werden.
fehlende Messwerte	Problem z. B. bei Metaanalysen, wo man ggf. geschätzte Werte einsetzt, um weitere Berechnungen möglich zu machen
Störgröße, Störfaktor (confounder)	> Parameter oder Risikofaktor, der sowohl mit der Exposition / Intervention als auch mit der Zielgröße (z. B. Erkrankung) assoziiert ist. Bsp.: Alter, Geschlecht, Nikotinkonsum, zusätzliche Erkankungen > Abhilfe: im Studiendesign z. B. Randomisierung, in Datenanalyse z. B. Stratifizierung
Schwankungen im Krankheitsverlauf	> Besserung oder Verschlechterung unabhängig von Intervention, z. B. spontane Besserung von Erkältungskrankheiten, Fortschreiten eines Diabetes, Schübe bei Multipler Sklerose oder Rheuma > Abhilfe: Kontrollgruppe
Selektionsbias (selection bias)	> systematische Unterschiede in der Auswahl der Probanden oder der Zusammensetzung der Gruppen, z. B. Männer vs. Frauen, Patienten mit leichten vs. schweren Symptomen > Abhilfe: Randomisierung, verdeckte Zuordnung
Behandlerbias (performance bias)	systematische Unterschiede zwischen den Gruppen bezüglich z. B. Pflege, Aufmerksamkeit etc. > Abhilfe: Verblindung der Probanden und Behandler
Erinnerungsbias (recall bias)	erinnerungsbedingte Verzerrung, z. B. in retrospektiven Studien: Probanden, deren Erinnerung an eine mögliche Exposition ungenau ist, geben sie eher an, wenn sie erkrankt sind als wenn nicht.
Beobachterbias (observer / ascertainment bias)	systematische Unterschiede in der Bewertung, weil die Gruppenzuordnung nicht (ausreichend) verblindet ist
Informationsbias (detection bias)	systematische Unterschiede zwischen den Gruppen in der Bestimmung des Outcomes, z. B.: Hormonbehandlung bei Frauen führt zu mehr Arztbesuchen (z. B. wegen Blutungen), wodurch auch mehr Gebärmutter-Karzinome gefunden werden – obwohl Hormone keine Karzinome begünstigen.
Abweichungen vom Studienplan (attrition bias, drop out, loss to follow up)	systematische Unterschiede zwischen den Gruppen bei Abweichungen vom Studienprotokoll, z. B. wenn in der Interventionsgruppe mehr Probanden die Studie wegen Nebenwirkungen abbrechen als in der Plazebogruppe > Abhilfe: Intention-to-treat-Analyse
Publikationsbias (publication / reporting bias)	systematische Unterschiede zwischen veröffentlichten und unveröffentlichten Ergebnissen, weil z. B. signifikante, erwünschte und spektakuläre Ergebnisse häufiger, schneller und hochrangiger publiziert werden als andere

suchende Intervention, z. B. ein Medikament. Eine Erweiterung ist das sog. Cross-over-Design: Hier werden Interventions- und Kontrollgruppe oder 2 Interventionen zur Studienmitte gewechselt. Jeder Patient erhält also nacheinander Behandlung 1 und Behandlung 2.

> Verblindung: Wenn möglich, sollten die Beteiligten zusätzlich verblindet sein, d. h. die Probanden – und evtl. auch die Behandler und Auswerter (doppelte Verblindung) – wissen nicht, wer welcher Gruppe zugeordnet ist. Natürlich lässt sich das nicht immer erfüllen, z. B. wenn nur eine Gruppe operiert wird.

Dieses Studiendesign trägt dazu bei, dass beobachtete Unterschiede im Outcome zwischen den Gruppen mit größerer Sicherheit auf die untersuchte Intervention zurückgeführt werden können [3]. „Hier sollten Sie allerdings genau auf den untersuchten Endpunkt achten", empfiehlt der Experte Hans-Hermann Dubben. „Wenn der primäre Endpunkt die Blutdrucksenkung war und Ihnen das reicht, okay. Wenn Sie aber wissen wollen, ob das Medikament das Leben verlängert, muss beim primären Endpunkt etwas stehen zu Mortalität, Lebensdauer, Lebensqualität o. ä."

Abweichungen vom Protokoll Ein Nachteil des sehr „strengen" Studiendesigns der RCT ist, dass nicht alle Patienten das vorgesehene Studienprotokoll einhalten: Sie brechen die Behandlung z. B. wegen Nebenwirkungen ab, wechseln in eine andere Studiengruppe, versterben vorzeitig oder sind bei längeren Verlaufsstudien nicht mehr erreichbar („drop out", „loss to follow up"). Daraus kann ein systematischer Fehler entstehen, der sog. attrition bias (►Tab. 2). Dieser Punkt ist auch Dubben sehr wichtig: „Ich versuche immer nachzurechnen: Wurden tatsächlich alle Patienten ausgewertet? Wie viele sind verschwunden – und wie viele machen den gemessenen Effekt aus? Wenn das 40 Personen von 2000 sind, aber 250 Probanden verloren gingen, schaue ich genau, ob die aus beiden Gruppen gleichermaßen rausgefallen sind – sonst ist das Ergebnis zumindest zweifelhaft." Eine gewisse Abhilfe gegen Verzerrung bietet die Unterscheidung zwischen einer

> Intention-to-treat-Analyse (ITT-Analyse) und einer

> Per-Protocol-Analyse (PP-Analyse).

Intention-to-treat-Analyse Die ITT-Analyse berücksichtigt alle randomisierten Probanden in der ursprünglich geplanten Gruppenzuteilung – unabhängig davon, ob sie die Behandlung tatsächlich absolviert haben [2, 4]. Diese Auswertung ist „konservativ", d. h. ein Effekt in der Interventionsgruppe wird tendenziell unterschätzt [2, 4].

Per-Protocol-Analyse Die PP-Analyse dagegen berücksichtigt nur die Teilnehmer, die die Studie gemäß Studienplan beendet haben [4]. Sie erlaubt sozusagen die Abschätzung des Wirkungs-

potenzials unter optimalen Bedingungen [5] – mit dem Nachteil, dass die Randomisierung durchbrochen wird, die Studiengruppen evtl. nicht mehr vergleichbar sind, und man verzerrte Ergebnisse erhält. PP-Analysen überschätzen tendenziell den Effekt in der Interventionsgruppe und unterschätzen z. B. Nebenwirkungen [2, 4]. Eine PP-Analyse kann die ITT-Analyse nicht ersetzen, aber evtl. sinnvoll ergänzen.

Randomisiert-kontrollierte Studien (RCTs) mit einfacher oder doppelter Verblindung sind notwendig, um Ursache-Wirkungs-Beziehungen mit ausreichender Sicherheit nachzuweisen. Dafür muss die Studie alle Prinzipien aber auch tatsächlich einhalten.

„Und auch die beste RCT sollte Sie nicht dazu bringen, von heute auf morgen Ihre ärztliche Behandlungen zu ändern", warnt Dr. Dubben. „Man kann sich in seiner Meinung verunsichern lassen, aber muss eine Gesamtschau machen. Erkenntnisse aus Metaanalysen oder die gesammelte Expertise in Leitlinien werden selten durch eine einzige neue Studie umgeworfen." Am sichersten fühlt sich Dubben, wenn neben RCTs weitere Studien mit insgesamt konsistent interpretierbaren Ergebnissen zu einer Frage vorliegen, z. B. Laborergebnisse, Beobachtungs- und Fall-Kontroll-Studien.

Fall-Kontroll-Studie

Bekanntes Outcome Eine Fall-Kontroll-Studie geht von einem bekannten Outcome / Ereignis aus, z. B. dem Auftreten einer Krankheit oder einer Komplikation. Sie vergleicht dann die Probanden, bei denen dies eintritt („Fälle"), mit einer Gruppe ohne dieses Outcome („Kontrollen") [6]. Die Studie soll klären, ob die Gruppen sich in bestimmten Merkmalen unterscheiden, z. B.:

> War eine der Gruppen einem Risikofaktor stärker ausgesetzt?
> Hat eine der Gruppen eine Intervention häufiger erhalten?

Problem: unbekannte Störgrößen Da Fall-Kontroll-Studien immer retrospektiv angelegt sind, können sie keine Ursache-Wirkungs-Beziehung beweisen – es kann sich bei den beobachteten Zusammenhängen auch um zufällige Korrelationen handeln, oder es gibt zusätzliche unbekannte Störgrößen, die sowohl Exposition als auch Outcome beeinflussen (▶ Tab. 2). Auch Angaben zur Gesamt-Inzidenz einer Erkrankung bzw. eines Outcomes lassen sich aus Fall-Kontroll-Studien nicht gewinnen, da man ja die „Fälle" gezielt aussucht [7]. In manchen Fällen ist diese Art von Studien aber die einzig mögliche, z. B. wenn das betrachtete Outcome sehr selten ist.

Tab. 3 Auswertung einer epidemiologischen Fall-Kontroll-Studie (Beispiel).

Risikofaktor	Outcome	
	Patienten mit Herzinfarkt	Patienten ohne Herzinfarkt
Nichtraucher	29	3700
Raucher	77	950
Berechnung des Odds Ratio (OR)		
OR für Herzinfarkt bei Rauchern vs. Nichtrauchern: (77/950)/(29/3700) = 10,3 Das heißt: Patienten mit Herzinfarkt waren in dieser Studie mit einer 10,3-fachen Wahrscheinlichkeit Raucher statt Nichtraucher. Wie verlässlich das Ergebnis ist (statistische Signifikanz), ist damit aber noch nicht gesagt – hierfür muss man mit einem geeigneten statistischen Test den p-Wert und / oder das Konfidenzintervall berechnen.		

Eine Fall-Kontroll-Studie ist umso aussagekräftiger, je ähnlicher sich die Gruppen sind (außer in dem untersuchten Parameter natürlich). Um dies zu erreichen, kann man die Probanden zusätzlich „matchen", d. h. man wählt die „Kontrollen" so aus, dass sie z. B. in Alter, Geschlecht, Body-Mass-Index o. ä. den „Fällen" möglichst ähnlich sind [6, 7].

Kreuztabelle und Odds Ratio Die Ergebnisse von Fall-Kontroll-Studien werden oft in einer sog. Vierfeldertafel oder Kreuztabelle dargestellt (▶ Tab. 3). Das korrekte Effektmaß – also die Maßzahl, mit der das Ergebnis dargestellt wird – ist die „Odds Ratio" (OR, Chancen- oder Quotenverhältnis). Sie gibt an, um wie viel wahrscheinlicher z. B. das Outcome bei Vorliegen des Risikofaktors ist im Vergleich zur Gruppe ohne Risikofaktor. Ein OR von 2 bedeutet z. B. ein doppeltes Risiko, ein OR von 0,5 ein halbiertes.

Fall-Kontroll-Studien beweisen keine Kausalität. Sie liefern aber Vermutungen und Hypothesen, die sich ggf. in prospektiven Studien überprüfen lassen.

Metaanalyse

Zusammenschau von Einzelstudien Eine Metaanalyse hat das Ziel, möglichst die gesamte vorhandene Evidenz zu einer Frage zusammenzufassen. Sie führt die Ergebnisse mehrerer Einzelstudien – am besten mit gleicher Fragestellung und ähnlicher Methodik – statistisch zusammen, um durch die größeren Stichprobenumfänge verlässlichere Daten zu erhalten [5, 8, 9]. Viele Metaanalysen verwenden die Ergebnisse der Einzelstudien (meist gewichtet z. B. nach Fallzahl oder Variabilität), manche greifen aber auch auf die Originaldaten der Probanden zurück und werten sie erneut aus (gepoolte Auswertungen, Analysen mit Individualdaten) [9, 10]. Die Ergebnisse werden meist als gepooltes Effektmaß angegeben, z. B. als gepooltes Odds Ratio oder gepoolte / gewichtete mittlere

Bildnachweis: Matthias Enter / www.fotolia.com

Differenz. Das Studiendesign sollte folgende Punkte beachten [9]:

- definierte Such- und Einschlusskriterien für Einzelstudien
- Einbeziehung unpublizierter Literatur
- Qualitätsbewertung der Einzelstudien, nur Studien einer bestimmten Qualität sollten einbezogen werden [9]
- vorab Festlegung des Zielkriteriums und des Auswertungsverfahrens

Sind die Einzelstudien sehr unterschiedlich, ist nicht gesagt, dass man ihre Ergebnisse so einfach zusammenfassen kann. Verschiedene statistische Verfahren sollen dies berücksichtigen, z. B. Modelle mit festen vs. zufälligen Effekten oder Sensitivitäts- und Heterogenitätsanalysen.

Feste vs. zufällige Effekte Das Analyse-Modell mit festen Effekten nimmt an, dass

- die Einzelstudien den gleichen Effekt schätzen,
- Unterschiede nur durch zufällige Abweichungen zustande kommen und daher
- dasselbe Ergebnis herauskäme (etwa die gleiche Verbesserung durch ein Medikament), wenn z. B. die Stichproben unendlich groß wären.

Diese Annahme ist aufgrund der unterschiedlichen Versuchsbedingungen aber oft unrealistisch, sodass man für die Zusammenführung der Einzelergebnisse (d. h. Berechnung des gepoolten Effektschätzers) bei größerer Heterogenität ein Modell mit zufälligen Effekten wählt [10]: Dies berücksichtigt die Variabilität zwischen den Einzelstudien, daher ist das Gesamtergebnis tendenziell weniger eindeutig und z. B. die Konfidenzintervalle (in denen man den „wahren" Effekt vermutet) breiter [9].

Sensitivitäts- und Heterogenitätsanalysen Mit Sensitivitätsanalysen überprüft man, wie robust die Ergebnisse der Metaanalyse sind: Man wiederholt die Analyse z. B. mit anderen Einschlusskriterien oder anderen Annahmen für fehlende Werte und prüft, ob das zu einem abweichenden Gesamtergebnis führt. Heterogenitätstests prüfen, ob die Unterschiede zwischen den Einzelstudien größer sind als zufallsbedingt zu erwarten. Mögliche Ursachen für Heterogenität sind Charakteristika der Patienten, Unterschiede in den Interventionen oder Endpunkten [9, 10, 12]. Sind die Einzelstudien zu heterogen, kann es sinnvoll sein, sie gar nicht zusammenzufassen.

Publikationsbias Metaanalysen sind besonders vom Problem des sog. Publikationsbias (► Tab. 2) betroffen: Die veröffentlichten Einzelstudien sind evtl. nicht repräsentativ, denn „ungünstige", „unpassende" oder „langweilige" Ergebnisse sind für Wissenschaftler, Unternehmen und Fachzeitschriften weniger interessant, sie werden seltener, später und in weniger renommierten Zeitschriften publiziert [13, 14]. Keine neue Erkenntnis ist auch, dass industriefinanzierte Studien seltener publiziert werden als andere – und im Fall einer Veröffentlichung häufiger zu „positiven" Ergebnissen kommen [14]. Ein weiteres Problem von Metaanalysen sind mehrfache Veröffentlichungen einer einzigen Studie (v. a. bei erwünschten Ergebnissen): Große, multizentrische Studien werden z. B. in mehrere kleine unterteilt, oder die gleiche Untersuchung wird von verschiedenen Autoren in verschiedenen Zeitschriften „verwertet" [14].

Eine gewisse Abhilfe gegen den Publikationsbias sollen Studienregister bieten: Hier werden alle Studien zu ihrem Beginn registriert und Änderungen z. B. im Studienprotokoll dokumentiert. „Das ist eindeutig ein Fortschritt", so Dubben. „Er zeigt sich allerdings erst allmählich, indem sich z. B. die Zeitschriften mehr und mehr leisten können, nur noch registrierte Studien anzunehmen." Einen Gegentrend sieht er allerdings in der stärkeren Abhängigkeit der Universitäten von Drittmittelprojekten: „Man ist nicht wirklich unabhängig, wenn man von der Firma auch in Zukunft Finanzierung braucht. Zumindest muss aber die Datenhoheit bei der Uni liegen – und natürlich muss alles veröffentlicht werden."

Metaanalysen fassen Einzelstudien zu einem Thema systematisch zusammen. Ihre Qualität hängt entscheidend von der Güte und Auswahl der Einzelstudien ab.

Worauf Sie im Detail achten sollten

Die Frage der Studie ist relevant für Sie, und die Herangehensweise scheint angemessen? Dann sollten Sie die Veröffentlichung genauer durchsehen. Sie müssen dabei nicht jedes Wort lesen – sollten aber wissen, wo die entscheidenden Dinge stehen. Auch ohne Statistik-Freak oder Experte im Fachgebiet zu sein, kann man häufig die folgenden Punkte beurteilen.

Autoren, Interessenkonflikte und Finanzierung Alle Autoren sollten mit ihrer Zugehörigkeit/Institution genannt werden, sodass ersichtlich ist, wenn z. B. Firmen beteiligt sind. Angaben zu Interessenkonflikten sind inzwischen vielfach vorgeschrieben – oft sind sie allerdings recht pauschal formuliert und beziehen sich auch nur auf die letzten Jahre bzw. das jeweilige Thema.
Teilweise wird auch die Finanzierung der Studie angegeben, z. B. durch eine Firma, öffentliche Institution oder Stiftung. Ob dieses Sponsoring einen ungebührlichen Einfluss auf die Ergebnisse zur Folge hat, ist für den Leser aber meist schwer zu beurteilen. Dr. Dubben wird v. a. misstrauisch, wenn eine Studie unnötig viele Standorte einbezieht: „Manchmal sind das Hunderte Studienzentren mit nur jeweils einem Dutzend Patienten – bei durchaus häufigen Krankheiten! Da kann der Haupt-Studienleiter viel leichter den Arzt vor Ort überreden, einzelne Patienten rauszunehmen, die ihm nicht passen."
Achten Sie auch auf „Danksagungen": Dies kann sich auf finanzielle Hilfe beziehen, aber auch auf „redaktionelle Unterstützung" o. ä. – wohinter wiederum ein Unternehmen stehen kann.

Interessenkonflikte oder Sponsoring bedeuten nicht gleich, dass man den Ergebnissen nicht trauen kann – sie sind aber evtl. mit erhöhter Vorsicht zu genießen.

Nun geht es an den eigentlichen Text. Jede Veröffentlichung einer Originalstudie sollte die folgenden Elemente enthalten.

Einleitung In der Einleitung sollten die Autoren den aktuellen Kenntnisstand skizzieren und darlegen, was sie mit ihrer Studie herausfinden wollten – und warum das wichtig ist [11]. Evtl. werden die verwendeten Modelle, Medikamente etc. vorgestellt. Als Leser sollten Sie erkennen können, ob die Fragestellung (für Sie) relevant ist und ob die Versuchsplanung ganz grob dem Zweck der Studie entspricht bzw. was die Grenzen der Studie sind [5, 8]. Außerdem sollte erwähnt sein, um was für eine Art von Studie (s. oben) es sich handelt.

Methoden Der Methoden-Teil ist quasi das „Kochrezept", anhand dessen andere Wissenschaftler das Experiment nachvollziehen und im Idealfall sogar nachmachen können. Folgendes sollten die Autoren – sofern gemäß Studiendesign möglich – schildern:

- Rekrutierung der Probanden (Charakteristika, Zahl, Zeitraum, Ort)
- Ein- und Ausschlusskriterien der Probanden (bei Metaanalysen: der Einzelstudien)

Fallzahlplanung

In der veröffentlichten Studie ist meist nicht erklärt, wie die Zahl der Probanden zustande kommt. Tatsächlich ist dies eine der entscheidenden Fragen, die vor Beginn der Studie zu klären sind. Da man aus ethischen sowie Zeit- und Kostengründen nicht unnötig viele Teilnehmer rekrutieren möchte [8, 11], versucht man, vorher abzuschätzen:

- Welcher Verteilung folgen die relevanten Maßzahlen (Normalverteilung o. a.)?
- Welchen Effekt erwartet man – und in welcher Größe?
- Wie viele Probanden braucht man mindestens (pro Gruppe und insgesamt), um diesen Effekt statistisch signifikant nachweisen zu können?
- Wie viele muss man anfragen oder rekrutieren, um – trotz ungeplanter Ausfälle – die notwendige Zahl zu erreichen?

- Einverständnis der Ethikkommission und der Teilnehmer
- ggf. Registrierungsnummer der Studie (z. B. bei www.clinicaltrialsregister.eu oder http://clinicaltrials.gov)
- ggf. Randomisierungsverfahren, Verblindung
- zeitlicher Ablauf der Studie
- genaue Behandlung der Interventions- und Kontrollgruppe
- ggf. Erfassung von Nebenwirkungen
- gemessene Parameter, Messmethoden, verwendete Geräte, Medikamente etc.
- Endpunkte („outcome"), ggf. primäre und sekundäre
- statistische Auswertungsmethoden

Als Leser kann man die folgenden kritischen Fragen stellen:

- Waren die Endpunkte vorab klar definiert? Sonst können die statistischen Analysen keine Antworten liefern, sondern nur aufgrund von Beobachtungen Hypothesen generieren [15].
- Sind die untersuchten Probanden repräsentativ für die Patienten bzw. die Bevölkerung? Wurde für relevante Störgrößen (Confounder) kontrolliert? [5]
- Ist die Verblindung sicher – bei Probanden und ggf. Behandlern [15]?
- Waren Untersuchungsmethode und -zeitpunkt geeignet in Bezug auf Wirkeintritt, Therapieziel etc. [5]?
- Ist der Vergleich fair – ist z. B. die Standardtherapie richtig dosiert und die neuartige Behandlung im Alltag praktikabel [5, 15]? Ist die Standardtherapie zu niedrig dosiert, überschätzt man die Wirkung des neuen Präparats, ist sie zu hoch dosiert, verstärkt dies die Nebenwirkungen [4].
- Sind die untersuchten Endpunkte klinisch relevant? Maßzahlen wie Blutdruck, HbA_{1c}-Wert oder Knochendichte sind als „Surrogatparameter" zwar leicht zu messen und evtl. auch leicht zu beeinflussen – haben aber evtl. keine relevante Wirkung auf Morbidität, Mortalität o. ä. [4, 5, 15].
- Bei Metaanalysen: Sind die Kriterien für Ein- oder Ausschluss der Einzelstudien festgelegt? Ist die Suchmethode geeignet, um alle durchgeführten Studien zu finden? Wurde die Qualität der Einzelstudien bewertet, sind z. B. alle randomisiert [8]? Oder wurden Studien willkürlich ausgeschlossen [16]?

Bildnachweis: iStockphoto

Ergebnisse Hier stellen die Autoren ihre Resultate dar – möglichst übersichtlich und wertfrei, bezogen auf die vorher genannten Endpunkte. Evtl. ist der Abschnitt gegliedert in

- einen deskriptiven Teil („Beschreibung" der Daten, häufig anhand von Tabellen und Grafiken) und
- einen analytischen Teil (Schätzung der Effektstärken, Zusammenhänge zwischen Merkmalen, Wechselwirkungen zwischen Einflussfaktoren, Signifikanzen und Konfidenzintervalle) [11].

Hans-Hermann Dubben greift an dieser Stelle regelmäßig zum Bleistift und rechnet die wichtigsten Ergebnisse nach – soweit möglich. „Manchmal sind ja nur relative Häufigkeiten angegeben", bedauert er. „Herausgeber von Fachzeitschriften sollten darauf bestehen, dass die dazugehörigen absoluten Zahlen mit angegeben werden." Besonders interessiert ihn, wie viele Patienten den entscheidenden Unterschied ausmachen. „Ich will wissen: Wie viele mehr müssten z. B. in der einen Gruppe krank sein, damit der Effekt weg ist? Wenn das sehr wenige sind – oder deutlich weniger als die ausgeschlossenen oder ausgeschiedenen Patienten – sind die Ergebnisse kaum glaubwürdig. Egal, wie klein der p-Wert ist." Potenzielle Messfehler oder kleine Manipulationen können dann das Gesamtergebnis stark beeinflussen [17].

In einer aktuellen Veröffentlichung [18] diskutiert Dubben das Beispiel der bisher größten Studie zum Prostatakrebs-Screening, der „European Randomized Study of Screening for Prostate Cancer" [19]: Sie bescheinigte dem Screening eine – auf den ersten Blick beeindruckende – relative Senkung der Prostatakarzinom-Mortalität von 20 %. „Bei genauer Betrachtung waren es aber nur 61 Personen, die diesen Effekt ausmachten", so Dubben. „Von 180 000 Probanden insgesamt, von denen nochmal 27 000 aus wenig überzeugenden Gründen ausgeschlossen wurden. 61 als Zünglein an der Waage sind dann zu wenig."

Aber auch wenn man annimmt, dass die Daten korrekt erfasst und verrechnet wurden, unterscheiden sich bessere und schlechtere Studien z. B. in folgenden Fragen:

- Sind die primären und sekundären Endpunkte sinnvoll gewählt? Primäre Endpunkte sind die wichtigsten und müssen vor Studienbeginn definiert werden, sekundäre nutzt man, um z. B. zusätzliche Effekte der Intervention zu klären. Zusätzliche Subgruppenanalysen sollen ggf. den Effekt der Intervention auf eine oder mehrere Untergruppen von Probanden prüfen (z. B. getrennt nach Geschlecht oder Alter).
- Sind die gewählten Effektmaße / Maßzahlen sinnvoll und aussagekräftig? Eine relative Risikoreduktion von 30 % mag sich imposant anhören – wenn es dabei aber z. B. um die Prävention eines seltenen Ereignisses geht, ist die absolute Risikoreduktion (d. h. die Risikoreduktion in der Gesamtbevölkerung) entsprechend gering und die „Number needed to treat" entsprechend groß.
- Sind die Ergebnisse auch klinisch relevant? Eine Blutdrucksenkung um 1–2 mmHG oder eine Schmerzreduktion um 0,3 auf einer Skala von 1–10 sind z. B. klinisch wenig relevant – erst recht nicht, wenn dieser Fortschritt mit neuen Nebenwirkungen oder hohen Kosten erkauft wird [3, 4].
- Könnten einige Probanden oder Untersucher unzureichend verblindet gewesen sein?
- Haben Autoren fehlende Messwerte willkürlich ergänzt? [16]
- Falls die Studie vorzeitig abgebrochen wurde: Ist der Grund plausibel? War das Abbruchkriterium vorher definiert? [4]
- Bei Metaanalysen: Wurden keine Einzelstudien doppelt verwendet?

Diskussion Im Diskussions-Teil interpretieren die Autoren ihre Ergebnisse und vergleichen sie mit dem aktuellen Stand der Wissenschaft oder der üblichen Behandlung [8, 11]. Sind die Ergebnisse plausibel oder überraschend? Sind sie auf andere Fälle bzw. Patienten übertragbar [8]? Welche neuen Fragen ergeben sich? Hans-Hermann Dubben liest die Diskussion als letztes – wenn überhaupt. „Es kann ganz interessant sein, welche Botschaft die

Statistische Signifikanz

„Es gibt ein unglaubliches Missverständnis, was der p-Wert leistet", erläutert Hans-Hermann Dubben eines seiner Lieblingsthemen. „Wenn ein Effekt mit einem p-Wert von 5 % statistisch signifikant ist, sagt das so gut wie gar nichts!" Korrekt interpretiert bedeute er lediglich:

- Wenn beide Interventionen in Wirklichkeit gleichwertig sind, kann die beobachtete oder eine größere Differenz mit 5 % Wahrscheinlichkeit zufällig auftreten.

„Es kann sich um einen echten Effekt handeln – aber auch um einen Fehler im Studiendesign oder den Ausschluss von relevanten Patienten", so Dubben. „Oder man testet einfach so oft, bis man ein signifikantes Ergebnis hat!" Auch wenn in Wirklichkeit kein Unterschied besteht, wäre das bei p = 5 % im Schnitt jedes zwanzigste. Das heißt: Man kann z. B. Patientengruppen oder Zeiträume solange unterteilen und testen, bis man irgendwo ein signifikantes Ergebnis findet, wo in Wahrheit nur eine zufällige Korrelation besteht (z. B. Sternzeichen und OP-Ergebnis). Statistisch aussagekräftig ist das Ergebnis einer Subgruppenanalyse daher nur, wenn diese inhaltlich sinnvoll ist und von vornherein beim Entwurf der Studie eingeplant wurde – oder zumindest das Signifikanzniveau entsprechend angepasst wird, z. B. indem man einen kleineren p-Wert als Grenze für Signifikanz festsetzt.

Infoquellen im Internet

Buch „Wo ist der Beweis? Plädoyer für eine evidenzbasierte Medizin“, erhältlich im Volltext unter:
> http://de.testingtreatments.org/wp-content/uploads/2013/07/wo_ist_der_beweis_volltext.pdf

Klinische Studien als Zeichentrickfilm:
> www.ecranproject.eu/de/content/zeichentrickfilm-zu-klinischen-studien

CONSORT (Consolidated Standards of Reporting Trials) zum Veröffentlichen von RCTs:
> www.consort-statement.org (auch dt. Version)

Cochrane Handbook for Systematic Reviews of Interventions:
> http://handbook.cochrane.org

Skeptical Journal Club: How To Read A Medical Study
> www.youtube.com/watch?v=RJFH8sUSzI4

Dr Shaneyfelt‘s Approach to Reading a Clinical Research Study
> www.youtube.com/watch?v=Gz_gu7pPB7s

Fazit

Wirklich gute Studien ohne methodische Mängel sind selten. Beim Lesen sollten Sie daher die wichtigsten kritischen Fragen im Hinterkopf haben – und sich klar sein, was Sie von der Studie wissen wollen. Ob sie das überhaupt leisten kann, sagt einem oft schon das Studiendesign. Wenn Sie Mängel feststellen, können Sie die Studienergebnisse zwar trotzdem zur Kenntnis nehmen – sollten aber auf Bestätigung aus weiteren Quellen warten, bis Sie auf dieser Grundlage z. B. Ihre Therapie ändern.

Schauen Sie zum Schluss auch noch einmal ins Abstract: Sind dort wirklich die wesentlichen Ergebnisse der Studie zusammengefasst – oder suggeriert das Abstract (auch in Verbindung mit der Überschrift) evtl. Resultate, die so klar gar nicht erzielt wurden?

Autoren aus der Arbeit herauslesen“, meint er. „Jeder spielt ja in einer bestimmten Mannschaft oder gehört zu einer gewissen Schule – das erkennt man oft in der Diskussion.“ Auch Angaben zu den Schwächen der Arbeit interessieren ihn: „Die Autoren wissen schließlich am besten, welche Patienten nicht ausgewertet werden konnten oder welche Tests nicht funktioniert haben.“

Als Leser können Sie auch kritisch fragen:
- Sind die Erkenntnisse wirklich neu?
- Könnte man die Daten auch anders interpretieren – und so evtl. zu anderen Aussagen kommen?
- Sind die Schlussfolgerungen der Autoren wirklich durch die Daten gestützt, oder picken sie sich einseitig die gewünschten Ergebnisse heraus? Gehen sie korrekt mit nicht signifikanten Ergebnissen um?
- Sind die geprüften Produkte oder Verfahren alltagstauglich für Ärzte und Patienten? Oder handelt es sich um eine riskante, experimentelle oder sehr teure Behandlung, die in der Praxis evtl. (noch) gar nicht umsetzbar ist [5]?
- Sind Nebenwirkungen und unerwünschte Wirkungen der Behandlung angegeben [5]?
- Geben die Autoren Grenzen und mögliche Fehlerquellen an [11]?
- Identifizieren sie weiteren Forschungsbedarf [8]?

Manche Studien haben noch einen kleinen letzten Teil „Schlussfolgerungen“, in dem Fragestellung, wesentliche Resultate und ihre Interpretation zusammengefasst sind [11].

Literaturangaben Anhand der häufig langen Liste der Literaturangaben können Experten auf dem Fachgebiet nachvollziehen, ob die wichtige Literatur berücksichtigt wurde. Als Nicht-Spezialist kann man evtl. nur anhand der Jahreszahlen prüfen, ob auch aktuelle Literatur aufgeführt (und damit hoffentlich aktuelles Wissen einbezogen) ist.

„Trauen Sie der Statistik nicht zuviel Beweiskraft zu“, rät Dr. Dubben abschließend. „Fragen Sie sich: Wie soll das eigentlich biologisch oder physikalisch funktionieren, was da statistisch signifikant gezeigt wurde? Gibt es überzeugende Experimente oder Theorien zum Wirkmechanismus? Wenn die Gesamtschau von Theorie, Experiment, klinischen Studien und Statistik ein konsistentes Bild ergibt, dann ist man der Wahrheit näher als allein mit Statistik.“

Julia Rojahn

Literatur online Das Literaturverzeichnis zu diesem Beitrag finden Sie im Internet: Unter„www.thieme-connect.de/products“ können Abonnenten und Nichtabonnenten die Seite der Lege artis aufrufen und beim jeweiligen Artikel auf „Ergänzendes Material“ klicken – hier ist die Literatur frei zugänglich.

Weiterführende Literatur Statistik unterhaltsam mit Büchern von Hans-Hermann Dubben und Hans-Peter Beck-Bornholdt:
> Der Hund, der Eier legt – Erkennen von Fehlinformation durch Querdenken. 8. Aufl. Reinbek: Rowohlt; 2006
> Mit an Wahrscheinlichkeit grenzender Sicherheit – Logisches Denken und Zufall. 6. Aufl. Reinbek: Rowohlt; 2005

Workshops Cochrane Deutschland bietet Workshops u. a. zu systematischer Literaturrecherche und systematischen Übersichtsarbeiten an: www.cochrane.de/workshops

Fortsetzung im nächsten Heft In der nächsten Ausgabe der Lege artis lesen Sie in Teil 2 unserer Statistik-Serie das Wichtigste zu Konfidenzintervall, Hypothesentests, Signifikanzen und Co.

Beitrag online zu finden unter http://dx.doi.org/10.1055/s-0041-103663