

3. Qualitätsbewertung klinischer Studien (1)

Wie man die Spreu vom Weizen trennt

Birgit Schindler, Judith Günther und Katja Suter

Im vorliegenden Kapitel wird hergeleitet, warum die randomisierte kontrollierte Studie der Goldstandard für den Beweis einer Ursache-Wirkungs-Beziehung ist. Um Fehlerquellen bei der Studienbewertung zu identifizieren, ist es nötig, sich zunächst mit einigen Begriffen vertraut zu machen. Anhand einer Studie zu der Frage „Hilft Zink bei Erkältung?“ wird dann exemplarisch dargestellt, wie man bei einem „critical appraisal“ strukturiert vorgehen kann. Hierfür sind Checklisten hilfreich, die das Vorgehen bei einer kritischen Durchsicht systematisieren und gleichzeitig auch dokumentieren. Zunächst wird die Methodenqualität (interne Validität) der Studie geprüft. Die Fragen, die dahinter stecken, lauten: „Kann ich dem Ergebnis dieser Studie überhaupt trauen?“ und „Lohnt es sich weiter zu lesen?“.

In der Beratung rund um Arzneimittel gilt es, Fragen nach Nutzen und Schaden einer Therapie zu beantworten. Das Anliegen einer evidenzbasierten Pharmazie ist es, Beratungsentscheidungen eine wissenschaftlichere Grundlage zu geben, als dies oft der Fall ist. Ist Therapie A wirksamer als Therapie B? Wie sicher ist das? Worin liegen die Unterschiede zwischen den Therapien C und D? Worin unterscheidet sich das Nebenwirkungsspektrum der Therapien E und F? Um diese Fragen beantworten zu können, werden glaubwürdige und möglichst fehlerfreie Informationen zu diesen Therapien benötigt.

3.1. Der Goldstandard für einen Beweis: Die randomisierte kontrollierte Studie (RCT)

Randomisierte kontrollierte Studien (RCT) stellen die zuverlässigste Versuchsanordnung dar, um den Effekt von zwei Therapien miteinander zu vergleichen. Qualitativ hochwertige RCT mit geringem Risiko für systematische Fehler (Bias) haben neben hochwertigen systematischen Reviews oder Metaanalysen den höchsten Evidenzgrad, das heißt, sie liefern die höchste Sicherheit bezüglich der dargestellten Zusammenhänge (Tab. 3.1).

Randomisierung

Bei einer randomisierten kontrollierten Studie werden die Patienten mit einem allein vom Zufall abhängigen Verfahren durch Verwendung von Zufallszahlen oder Computeralgorithmen auf verschiedene Behandlungsgruppen verteilt (*Randomisierung*). Dadurch soll sichergestellt werden, dass alle bekannten, aber auch unbekanntem potenziellen Störgrößen

auf den Krankheitsverlauf, wie Risikofaktoren für Erkrankungen oder bestimmte Charaktereigenschaften der Patienten, in den Behandlungsgruppen gleich verteilt sind (Strukturgleichheit zwischen den Gruppen). Als Störgrößen – auch als „Confounder“ bezeichnet – können zum Beispiel Gewicht, Alter oder Geschlecht der Patienten, individuelles Gesundheitsverhalten wie Rauchen, körperliche Fitness oder der sozioökonomische Status Einfluss auf das Therapieansprechen nehmen. Nur bei *Strukturgleichheit der Behandlungsgruppen* kann später ein beobachteter Unterschied zwischen den Therapien tatsächlich auf den Einfluss der Behandlung zurückgeführt und nicht aufgrund des Einflusses von Confoundern verfälscht werden. Das Randomisierungsverfahren darf nicht vorhersehbar sein, damit von keinem der Beteiligten während der Studienrekrutierung die Zuteilung gewollt oder ungewollt beeinflusst werden kann oder von der Studienteilnahme aufgrund der in Aussicht stehenden Zuteilung ganz abgesehen wird [7]. Über den gesamten Studienzeitraum muss daher die Randomisierungsliste der Studienteilnehmer zu den Behandlungsgruppen allen direkt an der Behandlung beteiligten Personen verborgen bleiben („*concealment of allocation*“, Geheimhaltung der Behandlungszuteilung). Dies erfordert eine Trennung zwischen den Personen, die die Randomisierung vornehmen, und den Personen, die die Behandlung durchführen. Eine verdeckte Zuteilung ist nicht zu verwechseln mit der Verblindung, die im Folgenden beschrieben wird [8].

Aber auch bei einer randomisierten Zuteilung können allein durch Zufall Unterschiede zwischen den Gruppen entstehen, besonders wenn die Anzahl der randomisierten Patienten klein ist. Deshalb wird von den Studienautoren bei der Publikation nach CONSORT (siehe Kapitel 3.2) eine Tabelle

Tab. 3.1. Evidenzklassen [1]

Klasse	Anforderungen an die Studien
I	Ia Evidenz aufgrund einer systematischen Übersichtsarbeit randomisierter kontrollierter Studien (evtl. mit Metaanalyse)
	Ib Evidenz aufgrund mindestens einer qualitativ hochwertigen randomisierten kontrollierten Studie
II	Ila Evidenz aufgrund mindestens einer gut angelegten kontrollierten Studie ohne Randomisierung
	Ilb Evidenz aufgrund einer gut angelegten, quasi-experimentellen Studie
III	Evidenz aufgrund gut angelegter, nicht experimenteller deskriptiver Studien
IV	Evidenz aufgrund von Berichten/Meinungen von Expertenkreisen, Konsensuskonferenzen und/oder klinischen Erfahrungen anerkannter Autoritäten

gefordert („table one“), in der die Verteilung der für den Behandlungserfolg relevanten Merkmale in Interventions- und Kontrollgruppe dargestellt wird.

Verblindung

Als „kontrolliert“ gilt eine Studie, wenn die Ergebnisse der Gruppe, an denen ein neues Verfahren erprobt werden soll (Interventionsgruppe), mit denen einer Kontrollgruppe verglichen werden. Als Kontrollbehandlung kann ein Placebo, die Standardtherapie oder auch keine Therapie dienen. Ein Placebo ist eine therapeutische Intervention, die von der aktiven Therapie nach Aussehen, Farbe, Geschmack und Geruch nicht zu unterscheiden ist, aber keine spezifische Wirkung hat. Es ist gezeigt worden, dass die „Wirkung“ von Medikamenten teilweise dem Placebo-Effekt (also dem „Akt der Einnahme“ oder dem mit der Verordnung in Zusammenhang stehenden Behandlungsakt) zuzuschreiben ist [2].

Neben der zufälligen und nicht vorhersehbaren Zuteilung auf die Behandlungsgruppen (Randomisierung) trägt die *Verblindung* zur Vermeidung von systematischen Fehlern während der Studiendurchführung bei. Denn alle Beteiligten werden ihr Verhalten ändern, wenn sie erfahren, wer welche Therapie erhält.

Im Idealfall wissen weder die Studienteilnehmer noch die Studienärzte oder die betreuenden Pflegefachkräfte und auch nicht die auswertenden Personen, welcher Gruppe die Patienten zugeordnet wurden. Manchmal sind Auswerter und Behandler in einer Person vereint. Wenn in diesem Fall weder Patient noch Studienarzt über die Zuteilung informiert sind, handelt es sich um einen doppelblinden Studienansatz. Nur mit einem solchen Studiendesign kann verhindert werden, dass die Ergebnisse unbewusst verfälscht werden [9]. Studien werden oft als verblindet (einfachblind), doppelblind oder dreifachblind bezeichnet – wichtig ist, dass in der Publikation erkennbar wird, wer tatsächlich verblindet war und dies auch bis nach Abschluss der Untersuchung geblieben ist. Eine Studie ohne Verblindung wird als „offene Studie“ bezeichnet. Dass nicht randomisierte oder nicht verblindete Therapiestudien „schlechter als nutzlos“ sind, da sie Fehlinformationen generieren, hat eine Londoner Arbeitsgruppe am Beispiel von

Studien zur renalen Denervierung bei therapieresistenter arterieller Hypertonie gezeigt. In **Abbildung 3.1** ist der Zusammenhang zwischen der Qualität des Studiendesigns und des gemessenen Behandlungseffekts dargestellt.

Studienauswertung: ITT versus PP

Bei der statistischen Auswertung einer randomisierten kontrollierten Studie ist wichtig, dass alle teilnehmenden Patienten in die Auswertung einbezogen werden. Das primäre Auswertungskollektiv ist das sogenannte *Intention-to-treat(ITT)-Kollektiv*, bei dem *alle* Patienten in der Behandlungsgruppe ausgewertet werden, der sie randomisiert zugeteilt wurden [5]. Nur so können die Vorteile einer Randomisierung (Strukturgleichheit der Behandlungsgruppen) für die Auswertung beibehalten bleiben. Da das ITT-Kollektiv alle Patienten umfasst, für die eine Behandlung vorgesehen war, fließen auch Daten von Patienten in die Auswertung ein, die zum Beispiel die Studienbehandlung nicht erhielten, die Studie unterbrochen oder vorzeitig beendeten. Neben der Aufrechterhaltung der Strukturgleichheit ist ein weiterer Vorteil dieser Vorgehensweise, dass das Ergebnis eher darüber Auskunft gibt, mit welchem Effekt in der therapeutischen Praxis zu rechnen ist, in der ebenfalls nicht alle behandelten Patienten eine verordnete Arzneimitteltherapie wie indiziert anwenden. Der Nachteil dieser Auswertungsstrategie ist, dass der Unterschied zwischen den Behandlungsgruppen undeutlicher wird, wenn viele Patienten die Therapie nicht erhalten, für die sie randomisiert waren.

Eine weitere Analysestrategie ist die Auswertung basierend auf dem *Per-Protocol(PP)-Kollektiv*. Dabei werden Patienten von der Auswertung ausgeschlossen, deren Behandlung vom ursprünglichen Behandlungsplan abgewichen ist. Bei der PP-Analyse werden somit nur die Daten der Patienten ausgewertet, die die ihnen zugeordnete Behandlung über den gesamten

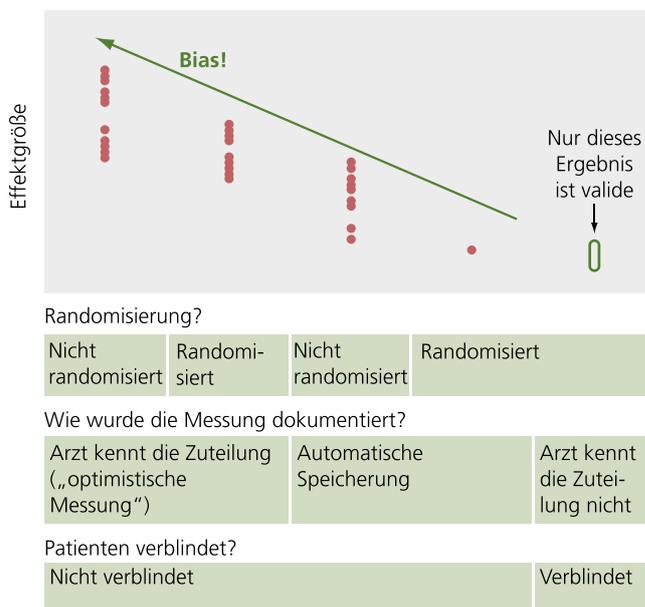


Abb. 3.1. Je höher die Qualität des Studiendesigns, desto geringer der gemessene Behandlungseffekt [mod. nach 11]

Studienzeitraum protokollgerecht angewendet haben. Das Ergebnis einer PP-Analyse gibt daher eine andere Auskunft als das der ITT-Analyse: Der mittels PP-Analyse errechnete Effekt einer Behandlung darf nur für die Patienten erwartet werden, die sich im therapeutischen Alltag vollständig an die Anweisungen des Arztes halten und therapietreu sind. Bei der PP-Analyse ergibt sich ein Risiko für Fehlinterpretationen, da mit dem Ausschluss von Patienten aus der Analyse die anfänglich durch die verdeckte Randomisierung erzielte Vergleichbarkeit der Gruppen (Strukturgleichheit) verloren gehen kann. Es besteht folglich die Gefahr, dass die ausgeglichene Verteilung der bekannten und unbekanntem Confounder aufgehoben wird und somit der beobachtete Unterschied nicht (oder nicht mehr allein) durch die unterschiedliche Behandlung, sondern auch durch andere, die Prognose beeinflussende Faktoren (Confounder) zustande gekommen ist.

3.2. Qualitätsmerkmale eines RCT: Wo lauern Fehlerquellen?

Um die Glaubwürdigkeit der Ergebnisse aus RCT (interne Validität; siehe **Infokasten 3.1**) einschätzen zu können, müssen Leser das Design, die Durchführung und die Analyse der Studie nachvollziehen können. Grundvoraussetzung für die Überprüfung der internen Validität ist, dass die Autoren ihre Methoden in der Studienpublikation vollständig transparent machen. Internationale Empfehlungen zur Berichterstattung randomisierter klinischer Studien finden sich im CONSORT-Statement („Consolidated standards of reporting trials“: Gemeinsame Standards für die Berichterstattung von Studien), das zum ersten Mal 1996 veröffentlicht und zuletzt 2010 überarbeitet wurde [10]. Inzwischen ist diese Leitlinie Teil der Autorenrichtlinien von über 70 internationalen Fachzeitschriften [12]. Das CONSORT-Statement besteht aus einer Checkliste und einem Flussdiagramm zur Beschreibung der Teilnehmerverluste im Verlauf eines RCT. Es kann von der CONSORT-Website (www.consort-statement.org) kostenlos heruntergeladen werden.

Im Folgenden wird – orientiert am CONSORT-Statement – dargestellt, wie man bei der Überprüfung der internen Validität einer Studie strukturiert vorgehen kann.

Systematische Fehler können im Studienverlauf an verschiedenen Stellen lauern (**Abb. 3.2**).

3.3. Warum es dennoch andere Studiendesigns geben muss

Gut durchgeführte randomisierte kontrollierte Studien liefern belastbare Daten für eine Ursache-Wirkungs-Beziehung. Denn durch methodische Verfahren wie Randomisierung, Verblindung und angemessene Analysemethoden sind die Ergebnisse solcher Studien vor systematischen Verzerrungen geschützt, die durch Differenzen zwischen den Behandlungsgruppen aufgrund bekannter und unbekannter Einflussfaktoren entstehen können [3]. RCT werden in aller Regel angewendet, um kausale Zusammenhänge für therapeutische, präventive und diagnostische Maßnahmen zu belegen. Nur

Infokasten 3.1

Interne Validität: Die interne Validität beschreibt, inwieweit die Ergebnisse einer Studie für die Stichprobe von Patienten, die untersucht wurde, korrekt sind. Bei der Überprüfung der internen Validität geht es darum, abzuschätzen, wie hoch das Risiko für die Verzerrung der präsentierten Ergebnisse ist.

Externe Validität: Die externe Validität beschreibt, inwieweit die Ergebnisse auf Patienten in der alltäglichen Praxis außerhalb der Studie übertragen werden können (Übertragbarkeit, Generalisierbarkeit).

sie erlauben eine sichere Abschätzung von Therapieeffekten, die man für Therapieentscheidungen benötigt.

Nicht immer sind RCT jedoch anwendbar. Bestimmte medizinische Fragestellungen, die beispielsweise die Prognose einer Erkrankung oder Risikofaktoren für eine Erkrankung betreffen, können nur durch Beobachten beantwortet werden. In bestimmten Fällen können Beobachtungsstudien auch zur Abschätzung von (unerwünschten) Effekten einer Arzneimitteltherapie von Bedeutung sein:

Fall-Kontroll-Studien

Mithilfe von *Fall-Kontroll-Studien* können beispielsweise seltene unerwünschte Ereignisse einer Arzneitherapie untersucht werden, die in einer randomisierten kontrollierten Studie aufgrund ihrer Seltenheit nicht und nur mit enormem Aufwand erfasst werden könnten. Dabei werden Personen mit dem unerwünschten Ereignis („Fälle“) mit „Kontrollen“

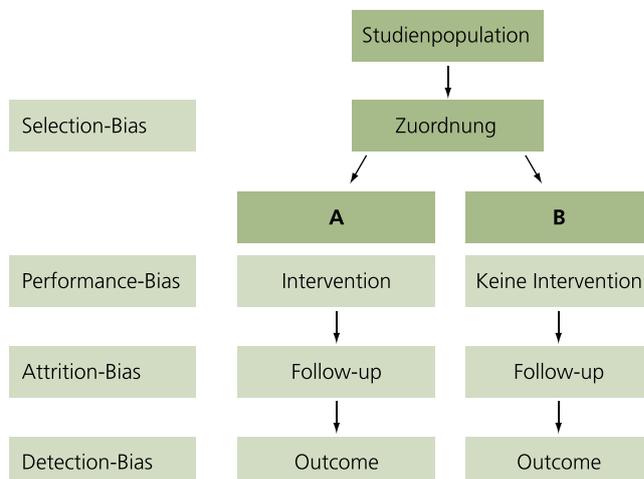


Abb. 3.2. Interne Validität: Überprüfung von Fehlerquellen

Selection-Bias: Entsteht durch eine ungewollte Selektion von Patienten bei Studienbeginn. Daher ist die verdeckte Randomisierung wichtig.

Performance-Bias: Entsteht, wenn bestimmte Patienten eine zusätzliche Behandlung oder eine engmaschigere Überwachung erhalten. Verblindung von Arzt und Patientenkollektiv minimieren diese Verzerrungsmöglichkeit.

Attrition-Bias: Entsteht durch das vorzeitige Ausscheiden von Studienteilnehmern aus einer Studie (z. B. Studienabbruch oder Follow-up-Verlust). Eine sorgfältige Studienplanung und Patientenaufklärung können den Follow-up-Verlust minimieren. Durch eine ITT-Analyse werden alle teilnehmenden Patienten in die Auswertung einbezogen.

Detection-Bias: Entsteht, wenn die Erfassung (Messung) nicht objektiv erfolgt oder von einem Endpunkt beeinflusst wird. Daher ist die Verblindung des Arztes entscheidend, der bei der Diagnose bzw. der Ermittlung der Zielgröße beteiligt ist.

verglichen, die dieses Ereignis nicht haben, aber ansonsten mit den „Fällen“ in definierten Merkmalen möglichst identisch sind.

Für dieses Untersuchungsdesign muss die Medikamentenanamnese eines jeden Falls und jeder Kontrolle mit Blick in die Vergangenheit rekonstruiert werden (= *retrospektiv*). Da es schwer ist, die Vergangenheit nicht aus der Sicht der gegenwärtigen Umstände eines Menschen zu interpretieren, liegt hierin Potenzial für einen systematischen Fehler [4].

Kohortenstudien

In *Kohortenstudien* werden Personen, die etwas gemeinsam haben (z. B. einen bestimmten Risikofaktor oder eine bestimmte therapeutische Intervention), zusammengestellt und mit Blick in die Zukunft (= *prospektiv*) hinsichtlich definierter Endpunkte beobachtet. Aus ethischen Gründen ist dieses Studiendesign zum Beispiel bei Schwangeren, die ein bestimmtes Arzneimittel einnehmen (z. B. Antiepileptika), mit vergleichbaren Schwangeren, die dies nicht tun, eine Möglichkeit, mehr über die Wirkungen oder unerwünschten Wirkungen von Arzneimitteln zu erfahren. Das Auftreten der definierten Endpunkte (z. B. Fehlbildungen bei den Nachkommen) in den beiden Gruppen wird prospektiv registriert.

3.4. Praxisbeispiel

Überprüfung der methodischen Qualität einer Studie

Ein Außendienstmitarbeiter war in Ihrer Apotheke und versuchte, Sie zu überzeugen, dass Sie das Zink-Präparat „Zink-Antiflu“ in die Sichtwahl aufnehmen sollten, da es ja Studien gäbe, die den eindeutigen Beweis lieferten, dass das Präparat bei Erkältung auch wirkt. Sie erinnern sich mit schlechtem Gewissen an ein Kundengespräch vor zwei Wochen, nach dem sie sich mit der Beleglage von Zink bei Erkältungskrankheiten beschäftigen wollten und sogar schon eine Studie ausgedruckt hatten. Leider blieb diese bislang auf Ihrem Schreibtisch unbearbeitet liegen. Bei genauerer Betrachtung des Firmenprospekts, das der Außendienstler zur Untermauerung seiner Behauptungen nutzte, stellen Sie fest, dass darin eben diese Studie zitiert wird, die auf Ihrem Schreibtisch auf Durchsicht harret. Sie beschließen, sich nun wirklich genauer damit zu beschäftigen, und fragen sich, wie denn die Vertrauenswürdigkeit einer solchen Studie überhaupt überprüft werden kann. Sie starten eine Google-Suche nach „quality of randomised controlled trials“. Rasch finden Sie verschiedene Checklisten, die alle in etwa gleich aufgebaut sind, beispielsweise vom CASP (Critical Appraisal Skills Programme) International Network. In **Tabelle 3.2** finden Sie die ausführliche Beurteilung der internen Validität am Beispiel der Studie von Prasad et al [6].

Lohnt es sich weiterzulesen? Ja, nachdem Sie die Checkliste bearbeitet haben, sind Sie der Meinung, dass die interne Validität der Studie recht gut ist. Die Studie scheint glaubwürdige Ergebnisse liefern zu können. Sie beschließen, für

heute Schluss zu machen und sich die Ergebnisse am nächsten Abend anzusehen.

Links zum Thema

- www.ebm-netzwerk.de/was-ist-ebm/grundbegriffe/glossar
Glossar des Deutschen Netzwerks Evidenzbasierte Medizin e. V.
- <http://www.casp-uk.net/casp-international>
Checkliste zur Bewertung der Studienqualität vom CASP (Critical Appraisal Skills Programme) International Network
- <http://www.consort-statement.org/>
Internationale Empfehlungen zur Berichterstattung randomisierter klinischer Studien durch das CONSORT-Statement („Consolidated standards of reporting trials“: Gemeinsame Standards für die Berichterstattung von Studien)

Literatur

1. Deutsches Netzwerk Evidenzbasierte Medizin e. V. <http://www.ebm-netzwerk.de/was-ist-ebm/images/evidenzklassen.jpg/view>
2. Fletcher, Robert H. Kontrollierte klinische Studien. In: Klinische Epidemiologie: Grundlagen und Anwendung/Fletcher/Fletcher/Wagner. Deutschsprachige Ausg. von Haerting J und Rink C. Wiesbaden: Ullstein Medical, 1999:193–214.
3. Fletcher, Robert H. Nachweis einer Ursache. In: Klinische Epidemiologie: Grundlagen und Anwendung/Fletcher/Fletcher/Wagner. Deutschsprachige Ausg. von Haerting J und Rink C. Wiesbaden: Ullstein Medical, 1999:330–44.
4. Fletcher, Robert H. Fall-Kontroll-Studien. In: Klinische Epidemiologie: Grundlagen und Anwendung/Fletcher/Fletcher/Wagner. Deutschsprachige Ausg. von Haerting J und Rink C. Wiesbaden: Ullstein Medical, 1999:297–305.
5. ICH-Guideline. E9 Statistical Principles for Clinical Trials, FDA Published in the Federal Register, 16. September 1998, Vol. 63, No. 179. Verfügbar unter: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf.
6. Prasad AS, Beck FW, Bao B, Snell D, et al. Duration and severity of symptoms and levels of plasma interleukin-1 receptor antagonist, soluble tumor necrosis factor receptor, and adhesion molecules in patients with common cold treated with zinc acetate. *J Infect Dis* 2008;197:795–802.
7. Schulz KF, Grimes DA. Reihe Epidemiologie 6: Generierung der Randomisierungslisten in randomisierten Studien: Zufall, nicht Auswahl. *Z Arztl Fortbild Qualitätssich.* 2007;101:419–426.
8. Schulz KF, Grimes DA. Reihe Epidemiologie 7: Geheimhaltung der Randomisierungslisten in randomisierten Studien: Wie man sich gegen Entschlüsselung wappnet. *Z Arztl Fortbild Qualitätssich.* 2007;101:499–505.
9. Schulz KF, Grimes DA. Reihe Epidemiologie 8: Verblindung in randomisierten Studien: Wie man verdeckt, wer was erhalten hat. *Z Arztl Fortbild Qualitätssich.* 2007;101:630–7.
10. Schulz KF, Altman DG, Moher D. CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c33210.
11. Shun-Shin M, Howard JP, Francis DP. Removing the hype from hypertension. Symplcity HTN-3 illustrates the importance of randomisation and blinding for exciting new treatments. *BMJ* 2014;348:g1937.
12. Türp JC, Antes G, Falck-Ytter Y. EBM-Splitter 15: Das CONSORT-Statement. <http://www.ebm-netzwerk.de/pdf/zahnsplitter/15.pdf>.

Tab. 3.2. Beurteilung der internen Validität einer Studie anhand eines Beispiels (Fortsetzung siehe nächste Seite) [6]

PICO-Frage: Kann eine Zink-Supplementierung eine Erkältung bei einem Erwachsenen verkürzen?	
Patienten: Erwachsene, die maximal seit 24 Stunden Erkältungssymptome hatten. Teilnehmer waren Medizinstudenten und Pflegepersonal des Detroit Medical Centers und Angestellte an der Wayne State University	
Verum: Zink-Lutschtablette (13,3 mg Zinkacetat), alle 2 bis 3 Stunden	
Kontrolle: Plazebo-Lutschtablette, alle 2 bis 3 Stunden	
Zielgröße: Durchschnittliche Dauer der Erkältungssymptome	
Fragen zur Methodenqualität und Aussagekraft (interne Validität)	
Strukturgleichheit	Kommentar für vorliegende Studie
Wurden die Patienten den Behandlungsgruppen rein zufällig zugeordnet (Randomisierung)?	Ja, eine ansonsten nicht an der Studiendurchführung beteiligte Person erzeugte den Randomisierungscode. 25 Patienten wurden der Plazebo-Gruppe zugeordnet und ebenfalls 25 der Zink-Gruppe
Wurde die Behandlungsfolge über den gesamten Studienzeitraum geheim gehalten („concealment of allocation“)?	Ja, die Verteilung der Medikation erfolgte durch einen verblindeten Studienassistenten (Trennung zwischen Zuteilung und Behandlung)
Hat die Randomisierung nachprüfbar funktioniert? Wurde hierfür eine Tabelle („table one“) vorgelegt, um die Vergleichbarkeit der Charakteristika zu Studienbeginn zu überprüfen?	In einer „Tabelle 1“ sind lediglich fünf Charakteristika dargestellt, in der Zink-Gruppe sind etwas weniger Raucher, dafür aber mehr Allergiker als in der Plazebo-Gruppe
Auswertungsgleichheit	Kommentar für vorliegende Studie
War die Messung des primären Endpunkts objektiv?	Nein, subjektiver primärer Endpunkt: Patienten dokumentierten täglich die Schwere von 10 Erkältungssymptomen auf einer vierstufigen Skala von 0 (= nicht vorhanden) bis 3 (=schwerwiegend)
Waren Patienten, Studienärzte und Auswerter gegenüber der Behandlungsart verblindet?	Studie wird als „doppelblind“ bezeichnet, Verblindung der Patienten und der Person, die die Studienmedikation verteilt, wird beschrieben
Wurde die Verblindung überprüft?	Plazebo-Lutschtablette enthält statt Zink entsprechend mehr Zucker, ansonsten identisch Verblindung wurde zu Beginn und am Ende der Studie überprüft: Nur zwischen 8% und 12% der Studienteilnehmer konnten ihre Gruppenzugehörigkeit richtig einschätzen, sodass die Verblindung trotz des schlechteren Geschmacks von Zink im Verum funktioniert zu haben scheint
Behandlungsgleichheit	Kommentar für vorliegende Studie
Wurden die Patienten – abgesehen von der zufällig zugeteilten Studienmedikation – in den jeweiligen Behandlungsgruppen ähnlich behandelt?	Es waren keine anderen Medikamente zur Behandlung der Erkältungssymptome erlaubt
Angaben zur Begleittherapie?	Die Patienten sollten alle Medikamente, die sie im Studienverlauf anwendeten, in ihrem Patiententagebuch dokumentieren. Über diese Dokumentation erfährt man jedoch nichts
Vollständige Nachbeobachtung	Kommentar für vorliegende Studie
Finden sich alle Patienten, die in die Studie aufgenommen wurden, in der Analyse wieder?	Es finden sich alle 50 Patienten in der tabellarischen Ergebnisdarstellung wieder
Wie viele Patienten beendeten die Studie vorzeitig und aus welchem Grund? Waren die Anzahl der ausgeschiedenen Patienten und die Gründe für das Ausscheiden zwischen den Behandlungsgruppen vergleichbar?	Es werden keine Angaben zum Loss-to-Follow-up gemacht. Die Studiendauer war mit maximal 8 Tagen (bzw. bis zum Ende der Symptome) kurz, sodass ein relevanter Loss-to-Follow-up auch nicht zu erwarten war. Ein Flussdiagramm – wie im CONSORT-Statement empfohlen – fehlt
Wurden alle Patienten in den ihnen zugewiesenen Behandlungsgruppen analysiert (Intention-to-treat, ITT)?	Ob alle 50 Patienten tatsächlich in der ihnen zugewiesenen Behandlungsgruppe ausgewertet wurden, ist unklar Die Analysemethode wird nicht erwähnt. Ob eine ITT-Analyse durchgeführt wurde, bleibt unsicher, ist aber anhand der Anzahl der ausgewerteten Patienten wahrscheinlich
Statistische Planung der Studie	Kommentar für vorliegende Studie
War die Teilnehmerzahl für die statistische Beweisführung ausreichend (Beschreibung der Fallzahlplanung)?	Eine nachvollziehbare Fallzahlberechnung ist angegeben

Tab. 3.2. Beurteilung der internen Validität einer Studie anhand eines Beispiels (Fortsetzung)

Fazit der methodischen Einschätzung

Insgesamt methodisch ordentlich.

Kleinere Einschränkungen aufgrund folgender Unsicherheiten:

- Waren die Gruppen hinsichtlich wichtiger Parameter ausreichend gut ausbalanciert? Dies ist nicht zuletzt aufgrund der geringen Teilnehmerzahl wichtig zu wissen, da bei kleinen Patientengruppen nicht in jedem Fall damit gerechnet werden kann, dass die Patienten durch die zufällige Zuteilung tatsächlich gleichmäßig auf die Behandlungsgruppen verteilt werden.
- In Tabelle 1, anhand derer dies abgeschätzt werden kann, sind insgesamt nur fünf Charakteristika dargestellt. In der Zink-Gruppe war die Zahl der Raucher mit $n=6$ (von insgesamt 25) etwas niedriger als in der Placebo-Gruppe ($n=9$ von insgesamt 25). Dagegen war die Zahl der Allergiker in der Zink-Gruppe ($n=6$) etwas höher als in der Placebo-Gruppe ($n=4$); dies deutet darauf hin, dass zufallsbedingt Unterschiede entstanden sein können, die Einfluss auf den Erkrankungsverlauf haben können. Über andere wichtige Parameter wie andere Atemwegserkrankungen in der Vorgeschichte, die Schwere der Symptome zu Studienbeginn, die Dauer der Symptome zu Studienbeginn (Ausschlusskriterium war eine Symptombdauer über 24 Stunden) oder auch die Art der Symptome (Einschlusskriterium war lediglich das Vorliegen von zwei der zehn Endpunkt-relevanten Symptome) erfährt man dagegen nichts. An anderer Stelle wird bei der Ergebnisdarstellung erwähnt, dass die Symptomschwere in der Placebo-Gruppe zu Beginn um 0,54 Punkte niedriger war als in der Zink-Gruppe (7,78 statt 8,32 von insgesamt 30 möglichen Punkten).
- Haben die Patienten noch andere Medikamente zur Linderung ihrer Symptome angewendet und gab es hier Unterschiede zwischen den Gruppen? Hierzu werden keine ausreichenden Angaben gemacht.
- Die Ergebnisdarstellung ist äußerst knapp gehalten. Es fehlen eindeutige Aussagen zur Analysemethode (ITT). Ein Flow-Chart – wie es auch im CONSORT-Statement gefordert wird – wäre wünschenswert.