Randomisierung und Vergleich der Ausgangswerte in klinischen Studien

Douglas G. Altman[1] und Caroline J. Doré[2]

[1] Medical Statistics Laboratory, Imperial Cancer Research Fund, London
[2] Section of Medical Statistics, Clinical Research Centre, Harrow, Middlesex, Großbritannien

aus: *Lancet* 1990; **335**: 149–53

Zusammenfassung

Überprüft wurden 80 Berichte über randomisierte klinische Studien aus vier führenden allgemeinmedizinischen Fachzeitschriften. Die Angaben zu den Randomisierungsverfahren waren unzureichend. In 30% der Studien war nicht eindeutig belegt, dass die Zuteilung zu den Gruppen randomisiert erfolgt war. In den Studien mit einfacher Randomisierung wiesen Behandlungs- und Kontrollgruppe allzu oft ähnliche Stichprobenumfänge auf, und wider Erwarten waren nur geringfügige Verzerrungen (Bias) zugunsten der geringeren Anzahl von Patienten in der experimentellen Gruppe zu beobachten. 41% der Studien enthielten keinen angemessenen Vergleich der Ausgangscharakteristika. Zur Verbesserung der Berichtsstandards werden Vorschläge unterbreitet.

Einleitung

Viele der in medizinischen Fachzeitschriften veröffentlichten Artikel enthalten statistische Fehler [1]. So waren beispielsweise in 86 kontrollierten Studien, über die in vier geburtshilflichen und pädiatrischen Zeitschriften berichtet wurde, nur 10% der Schlussfolgerungen gerechtfertigt, während in weiteren 71% die Angaben unzureichend waren [2]. In den klinischen Studienberichten aus vier allgemeinmedizinischen Fachzeitschriften wurden im Durchschnitt nur etwas mehr als die Hälfte (56%) von elf methodischen Aspekten beschrieben; vor allem das Randomisierungsverfahren wurde nur in einem Fünftel der Studien angegeben [3]. Eine solide durchgeführte randomisierte kontrollierte Studie ist der zuverlässigste Weg, um Therapien miteinander zu vergleichen. Zuverlässige Schlussfolgerungen sind aber nur dann gewährleistet, wenn beim Studiendesign und bei der Datenanalyse in mehrfacher Hinsicht sehr sorgfältig vorgegangen wird. Ohne eine Beschreibung der Studienmethodik müssen die Schlussfolgerungen einer Studie fragwürdig erscheinen. Die Randomisierung gewährleistet eine unverzerrte Zuteilung zu den Behandlungen, sie ergibt allerdings nicht zwangsläufig Gruppen, die sich in wichtigen prognostischen Faktoren ähnlich sind. Die Ähnlichkeit der Ausgangscharakteristika muss sichergestellt werden, aber nicht durch Hypothesentests [4–6]. Wir haben randomisierte klinische Studien aus vier allgemeinmedizinischen Fachzeitschriften überprüft und dabei die Verfahren der Behandlungszuteilung sowie die Darstellung und Interpretation der Ausgangswerte genauer untersucht. Da uns die Patientenzahlen in den Behandlungsgruppen oftmals allzu ähnlich vorkamen, haben wir die Gruppengrößen unter Berücksichtigung des Randomisierungsverfahrens untersucht.

Methoden

Aus den Zeitschriften *Annals of Internal Medicine, British Medical Journal, The Lancet* und *New England Journal of Medicine* wurden jeweils die ersten 20 randomisierten klinischen Studien ausgewählt, die nach dem 1. Januar 1987 erschienen sind. (Genauere Angaben zu den 80 identifizierten Studien sind auf Anfrage bei D. G. Altman erhältlich.) Diese Stichproben erstreckten sich über einen Zeitraum von 19, 13, 5 bzw. 10 Monaten. Unsere Untersuchung beschränkte sich auf Parallelgruppenstudien, in denen die Zuteilung zu zwei verschiedenen Behandlungen den Angaben zufolge randomisiert erfolgt war. Die erste Auswahl gründete sich auf das Abstract und die flüchtige Durchsicht des Volltextes. Einige Arbeiten wurden nachträglich ausgeschlossen, da sich nach gründlicher Lektüre zeigte, dass unsere Einschlusskriterien nicht erfüllt waren. Vor allem in einer als randomisiert bezeichneten Studie [7] war die Zuteilung nach ungeradem bzw. geradem Geburtsdatum erfolgt. Ferner wurden zwei Veröffentlichungen aus-

geschlossen, bei denen es sich nicht um die erste Publikation zu den betreffenden Studien handelte.

Die 80 Arbeiten wurden mit einem in einer kleinen Pilotstudie getesteten Standardauswertungsbogen untersucht. Beide Reviewer untersuchten unabhängig voneinander die Randomisierung und bewerteten die Gleichwertigkeit der Ausgangswerte, wobei abweichende Beurteilungen durch Diskussion geklärt wurden. Andere Aspekte wurden jeweils von nur einem Reviewer untersucht; dabei wurde darauf geachtet, dass derselbe in allen 80 Studien dieselben Themen bearbeitete. Im Hinblick auf die Art der Randomisierung bestanden zwischen beiden Reviewern kaum Diskrepanzen; die wenigen Unterschiede waren auf vage Angaben zur potenziellen Anwendung einer stratifizierten Randomisierung zurückzuführen. Bei der einfachen Randomisierung benutzt man eine einzige Folge von Zufallszahlen, um zu entscheiden, welche Behandlung ein Patient erhält, während bei der stratifizierten oder geschichteten Randomisierung für Subgruppen von Patienten – je nach prognostischen Faktoren oder Studienzentren – getrennte Zahlenfolgen verwendet werden. Bei beiden Randomisierungsverfahren kann man kleine Blöcke (einer festgelegten oder variablen Länge, z. B. jeweils sechs Patienten) bilden, um eine balancierte Zuteilung der Patienten zu den Behandlungsgruppen zu gewährleisten. Das Verfahren der Minimierung dient dazu, kleine, einander hinsichtlich mehrerer Charakteristika sehr ähnliche Gruppen zu bilden [8]. Zur Vermeidung von Bias sollte das Verfahren der Behandlungszuteilung vorher geplant werden: Zu den geeigneten Methoden gehören die zentrale Randomisierung, von der Apotheke vorbereitete kodierte Medikamente und die Verwendung von durchgehend nummerierten, undurchsichtigen und verschlossenen Umschlägen. Die Beurteilung der Frage, ob die Ähnlichkeit der Ausgangswerte angemessen bewertet worden war, erfolgte subjektiv und führte in 19 Fällen zu einer übereinstimmenden Bewertung. Zur Untersuchung der Anzahl der den Gruppen zugeteilten Patienten wurden die beiden Behandlungen anhand der verfügbaren Informationen als experimentell bzw. Kontrollbehandlung eingestuft.

Ergebnisse

Randomisierung

60% der Artikel (Tabelle 1) enthielten keine Angaben zum Randomisierungsverfahren. In einem Drittel der Studien war stratifiziert worden; nur eine Studie erwähnte eine einfache Randomisierung. In den meisten anderen Studien wurde wahrscheinlich eine einfache Randomisierung durchgeführt. In nahezu 30% der Studien wurde Blockbildung eingesetzt, darunter befanden sich aber nur 16 der 31 stratifizierten Studien (52%). Acht Studien (35%) enthielten keine Angaben zur Größe der Blöcke. Eine Studie mit 30 Patienten verwendete zu große (20er) Blöcke. Wird die Blockbildung ohne Stratifizierung durchgeführt, sollte die größte Differenz zwischen den Patientenzahlen in den beiden Gruppen nicht mehr als die Hälfte der Blockgröße betragen; in zwei Studien war dies nicht der Fall. Nur bei einer Studie führte man eine gewichtete Randomisierung durch, um eine ungleiche Aufteilung der Patientenzahlen zu erreichen. Aber auch in einer weiteren Studie lässt die Anzahl der den beiden Gruppen zugeteilten Patienten (40 und 78) ebenfalls auf eine Gewichtung schließen. Informationen über die zur Generierung von Zufallszahlen verwendete Methode stellen eindeutig unter Beweis, dass es sich um eine randomisierte Studie handelt. Entsprechende Angaben waren aber nur in 50% der untersuchten Studien zu finden: 16 Studien verwendeten Zufallszahlentafeln, 19 einen Computer, drei eine „zufällige Anordnung" und eine Studie das Verfahren der Minimierung. Fast die Hälfte der Studien (45%) machte keine Angaben über die Zuteilungsverfahren. Von den übrigen 44 Studien gaben nur 16 die Verwendung von Umschlägen an, von diesen wiederum erwähnten aber nur zwei, dass es sich dabei um nummerierte, verschlossene und undurchsichtige Umschläge handelte – allesamt wichtige Aspekte [9]. In vier Studien erfolgte die Zuteilung über eine zentrale Randomisierung. Wir gehen davon aus, dass in den 15 Studien, die über die Verwendung von nummerierten oder kodierten Flaschen berichteten, diese von einer Apotheke bereitgestellt wurden. Trotzdem berichteten nur 21 Studien (26%) über die Anwendung eines Verfahrens zur Reduktion von systematischen Fehlern (Bias). Nur 27 Studien (34%) enthielten sowohl Informationen über die zur Ge-

Tabelle 1. Randomisierung.

	Ann Intern Med	Br Med J	Lancet	N Eng J Med	Insgesamt
Art der Randomisierung					
einfach	0	0	1	0	1 (1%)
stratifiziert	12	5	4	10	31 (39%)
keine Angaben	8	15	15	10	48 (60%)
Block	6	6	4*	7	23 (29%)
stratifiziert und Block	5	3	3*	5	16 (20%)
Angaben zur Generierung der Zufallszahlen	10	10	7	12	39 (49%)
Bias verringernde Zuteilung					
ja	7	3	2	9	21 (26%)
nein	6	7	5	5	23 (29%)
keine Angaben	7	10	13	6	36 (45%)
Angaben zur Generierung der Zufallszahlen und Zuteilung	9	5	6	7	27 (34%)

Z. ärztl. Fortbild. Qual. Gesundh.wes. (2005) 99; 629–634
http://www.elsevier.de/zaefq

nerierung der Zufallszahlen verwendete Methode als auch über das Verfahren, nach dem die Patienten den Behandlungen zugeteilt wurden.

Stichprobenumfang

In 31 Studien (39%) beruhte der Stichprobenumfang auf vorher durchgeführten Berechnungen der statistischen Power (Tabelle 2). In weiteren 26% wurde die Studiendauer angegeben, obwohl nur selten erwähnt wurde, dass die Stichprobengröße von der Anzahl der Patienten abhängig war, die in einer zuvor festgelegten Zeitspanne rekrutiert worden waren. In einem Fall handelte es sich um eine sequenzielle Studie. Ein Drittel der Studien enthielt keine Angaben zu den verwendeten Fallzahlen. In 27 Studien (34%) wurden Patienten vor Beginn der Behandlung ausgeschlossen; nur neun Berichte enthielten genaue Angaben zur Anzahl der randomisierten Patienten, d. h. ein Viertel der 80 Studienberichte machte keine Aussage über die Anzahl der Patienten, die den einzelnen Behandlungen zu Beginn der Studie zugeteilt worden waren. In den neun Studien mit vollständigen Angaben wurden ca. 10% der Patienten vor Behandlungsbeginn von der Studie ausgeschlossen, und in 8 Fällen waren die Kontrollgruppen davon stärker betroffen ($p = 0{,}04$; Vorzeichentest). Die Mehrzahl der Studien, die über Studienausschlüsse berichtete, enthielt keine Angaben zur Anzahl der Ausschlüsse aus den einzelnen Gruppen, und in nur 7 Studien wurden diese Ausschlüsse auch begründet.

Von den 62 Studien, die Angaben zur Anzahl der randomisierten Patienten enthielten, wiesen 12 Studien pro Gruppe die gleiche Anzahl von Patienten auf; nur fünf davon hatten eine Blockrandomisierung durchgeführt. Von den 19 Studien mit Blockbildung hatten sieben Studien mehr Patienten in der Kontrollgruppe und sieben hatten mehr Patienten in der experimentellen Gruppe. Bei den 43 Studien ohne Blockbildung lag eine eindeutige Verzerrung vor: 26 (72%) der 36 Studien mit ungleichen Stichproben wiesen mehr Patienten in der Kontrollgruppe auf ($p = 0{,}01$; Vorzeichentest). Allerdings fiel der Unterschied in den Stichprobenumfängen im Allgemeinen eher gering aus (Abb. 1). 5% der Studien sollten außerhalb der beiden äußeren Geraden zu finden sein – keine der untersuchten Studien findet sich in diesem Bereich wieder. Weiterhin sollte sich die Hälfte der 43 Studien außerhalb des inneren Geradenpaares befinden – es sind aber nur fünf. Bei den 18 Studien, in denen nur die Anzahl der analysierten Patienten angegeben war, zeigte sich ein ähnliches Verteilungsmuster.

Tabelle 2. Stichprobengrößen und Ausschluss von Studienteilnehmern.

	Ann Intern Med	Br Med J	Lancet	N Eng J Med	Insgesamt
Angabe von Gründen für die verwendete Stichprobengröße					
Berechnung der Power	6	8	7	10	31 (39%)
Zeitraum	5	2	6	8	21 (26%)
gruppensequenzielles Verfahren	0	0	0	1	1 (1%)
keine Angaben	9	10	7	1	27 (34%)
Ausschluss von Studienteilnehmern					
kein Ausschluss	12	12	14	15	53 (66%)
wie randomisiert	3	4	2	0	9 (11%)
wie analysiert	5	4	4	5	18 (23%)

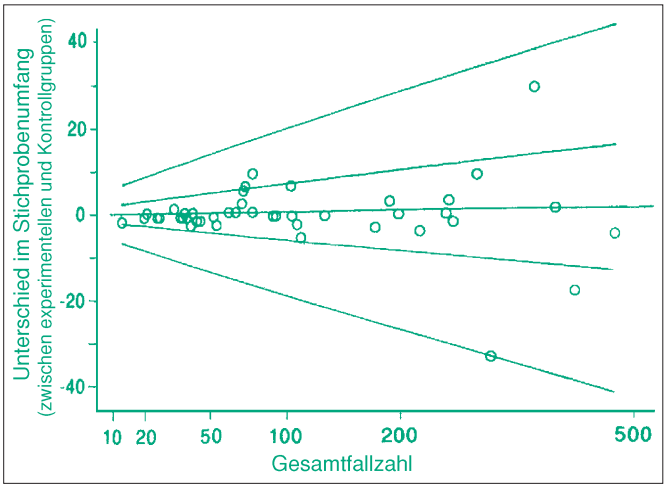

Abb 1. Beziehung zwischen unterschiedlichen Teilnehmerzahlen in experimentellen und Kontrollgruppen und der Gesamtstudiengröße in 43 Studien ohne Blockrandomisierung

Die Geraden geben die erwartete Verteilung an. Der Stichprobenumfang ist auf einer Quadratwurzelskala dargestellt, sodass die Konfidenzintervalle sich linear verbreitern. Die Breite des 95%-Konfidenzintervalls beträgt etwa $\pm 2\sqrt{n}$ (n Gesamtfallzahlstudiengröße).

Ausgangsmerkmale

Die Anzahl der Ausgangsmerkmale lag im Median bei 9, wobei 39% der Studien Daten für mehr als zehn Variablen angaben. Sechs Studien enthielten diesbezüglich keine Informationen. In 69 Arbeiten waren die Ausgangscharakteristika in Form kontinuierlicher Daten angegeben. In 19 Berichten wurde die Variabilität der Ausgangsdaten mit Hilfe des Standardfehlers (SE) und in einer Studie anhand von Konfidenzintervallen beschrieben. In 13 Artikeln

wurde kein Variabilitätsmaß angegeben. Damit war die Darstellung der Ausgangsdaten in 39 Studien (49%) unbefriedigend. In 46 Studien (58%) wurden für den Vergleich der Ausgangsvariablen Hypothesentests benutzt, aber in nur 34 Studien (74%) wurden auch die Methoden dargelegt. In 18 Studien (39%) wurden mehr als zehn Variablen getestet. In 17 dieser Studien (37%) bestand hinsichtlich einer oder mehrerer Ausgangsvariablen ein signifikanter Unterschied ($p < 0,05$) zwischen der experimentellen und der Kontrollgruppe. Insgesamt wurden in den 46 Studien 600 Hypothesentests durchgeführt, von denen 24 (4%) auf einem Niveau von 5% signifikant waren. Diese Werte beruhen auf den publizierten Analysen; Daten, für die die Autoren keine Testergebnisse vorgelegt hatten, wurden von uns nicht untersucht.

In nahezu 50% der Studien wurde nicht nach den Ausgangscharakteristika adjustiert, während ein Viertel ein statistisches Modell verwendete (Tabelle 3). Zwölf Studien untersuchten die Veränderungen gegenüber den Ausgangswerten. In acht Studien wurden die Analysen nur innerhalb der Behandlungsgruppen durchgeführt. Fast alle Studienberichte (91%) kommentierten in irgendeiner Form die Ähnlichkeit der Gruppen bei Studienbeginn. Wir bewerteten, inwieweit die Studienautoren mit diesen Vergleichen adäquat umgegangen waren. Dabei wurden die dargelegten Informationen, die Größe der Unterschiede zwischen den Gruppen, die Analysemethode sowie der Diskussionsumfang berücksichtigt. In 47 Studien (59%) wurde dieser Punkt angemessen behandelt: in zwanzig dieser Studien durch Modellbildung oder Untersuchung der Veränderungen gegenüber den Ausgangswerten, in weiteren 20 Studien durch adäquate Diskussion oder angemessenes Studiendesign. In sieben Studien waren die Gruppen einander so ähnlich, dass eine Diskussion nicht erforderlich war. Von den 33 Studien, die keine ausreichenden Vergleiche der Ausgangswerte enthielten, machten die meisten entweder ungenügende Angaben (17) oder hatten es versäumt, bei größeren Unterschieden zu adjustieren (13). „Größer" wird hier verstanden als beträchtlicher subjektiver Unterschied hinsichtlich der Mittelwerte oder Anteile unabhängig von der statistischen Signifikanz. Wir gingen davon aus, dass jede Ausgangsvariable potenziell prognostisch war. Drei Studien mit deutlich erkennbaren Unterschieden zwischen den Gruppen enthielten diesbezüglich – außer dem bloßen Hinweis auf das Vorliegen signifikanter Unterschiede – keinerlei Kommentar.

Tabelle 3. Handhabung der Ausgangswerte.

	Ann Intern Med	Br Med J	Lancet	N Eng J Med	Insgesamt
Methode					
Statistische Modellbildung	8	2	4	7	21 (26%)
Änderung gegenüber den Ausgangswerten	3	5	1	3	12 (15%)
Keine Anpassung	5	12	13	9	39 (49%)
Gruppeninterner Vergleich	4	1	2	1	8 (10%)
Angemessene Handhabung	15	12	8	12	47 (59%)

Diskussion

Randomisierung

Auch wenn in Leitlinien empfohlen wird, das Randomisierungsverfahren zu spezifizieren [10, 11], werden die Art der Randomisierung, die Quelle der Zufallszahlen und der Zuteilungsmechanismus im Allgemeinen nicht unterschieden, wobei Zelens Empfehlung [12] allerdings eine Ausnahme darstellt. Auch die Verblindung der Behandlungszuteilung ist wichtig, um sicherzustellen, dass die Studie frei von Bias ist [9]. Nur wenige Studienberichte enthielten – abgesehen von Hinweisen auf eine Stratifizierung – Angaben zur Art der Randomisierung. Eine Blockbildung wurde in 29% der Studien erwähnt, allerdings nur in 16 der 31 stratifizierten Studien. Ohne Blockbildung ergibt eine Stratifizierung keinen Sinn; allerdings ist es wahrscheinlich, dass die Blockbildung in mehr Studien erfolgte als angegeben. Über die Methoden zur Generierung der Zufallszahlen und den Mechanismus der Behandlungszuteilung wurden nur unzureichende Angaben gemacht – mehr als die Hälfte der Studien enthielt dazu keinerlei Informationen. Nahezu ein Drittel der Veröffentlichungen enthielt zu keinem der beiden Punkte Angaben, sodass der Nachweis fehlte, dass für die Studie randomisiert worden war. Bei 5–10% der „randomisierten" Studien stellte sich heraus, dass die Behandlungszuteilung nicht randomisiert erfolgt war [13, 14], sodass es sich bei einigen der 24 Studien, in denen entsprechende Angaben fehlten, sehr wohl um nichtrandomisierte Studien handeln könnte. Ein drastisches Beispiel für eine Verzerrung, die durch die systematische Behandlungszuteilung entstehen kann, beschreibt Keirse [15]. Ein Bias kann sich aber auch aus einer unverblindeten Behandlungszuteilung ergeben. In nur 26% der Studien wurde ein System verwendet, das geeignet war, systematische Fehler zu verringern.

Stichprobenumfang

Pocock et al. [16] stellten fest, dass die verwendete Stichprobengröße in nur fünf der im Jahre 1985 in drei allgemeinmedizinischen Fachzeitschriften erschienenen 45 Studien auf a priori durchgeführten Berechnungen der statistischen Studienpower beruhte. Bei unserer Untersuchung derselben Zeitschriften sowie einer weiteren 18 Monate später stellten wir fest, dass in 39% der 80 Studienberichte die Studienpower berechnet worden war. In den drei in beiden Reviews untersuchten Zeitschriften stieg diese Zahl von 11% auf 42% ($p = 0,001$; Chi-Quadrat-Test). Der von uns vorgelegte Wert (39%) ist der größte, der unseres Wissens jemals in einem Review ermittelt wurde, auch wenn dies immer noch be-

deutet, dass in ca. zwei Dritteln der Studienberichte keine Gründe für die Beendigung der Rekrutierung angegeben werden.

Außer in kleinen Studien gibt es keinen Grund für ähnliche Patientenzahlen in den Gruppen. Bei der einfachen Randomisierung kann es zu Diskrepanzen kommen, was aber auf die Studienpower keine schwerwiegenden Auswirkungen hat. Wir betrachteten die Verteilung der Größenunterschiede der Gruppen (wie randomisiert) unter Berücksichtigung des angegebenen Zuteilungsverfahrens. In den 19 Studien mit Blockbildung entsprachen die Unterschiede den Erwartungen. Bei den 43 Studien ohne Blockbildung waren die Stichprobengrößen in den beiden Gruppen einander allerdings viel zu ähnlich: Anstatt der erwarteten 50% lagen nur 5 Studien außerhalb der Innengeraden (siehe Abb.). Dieses Ergebnis stützt unsere frühere Hypothese. Die Clusterbildung um gleiche Stichprobenumfänge ist möglicherweise darauf zurückzuführen, dass die Studien keine Angaben zu folgenden Punkten machten: (A) Blockbildung, (B) Anwendung einer deterministischen Methode wie etwa die Zuteilung nach alternierendem oder geradem/ungeradem Datum oder (C) Korrektur eines unbefriedigenden Ungleichgewichts durch Zuweisung zusätzlicher Patienten zu einer Behandlung. Die Effektgröße lässt es unwahrscheinlich erscheinen, dass dies allein auf Zufall zurückzuführen ist. Wir glauben, dass sowohl (A) als auch (B) häufiger vorkommen, haben aber keinen Beweis dafür, dass die sehr viel bedenklichere Möglichkeit (C) auch tatsächlich angewendet wird.

Tendenziell war die Stichprobengröße in den Kontrollgruppen etwas höher als in den Behandlungsgruppen. Diese statistisch signifikante Asymmetrie, die sich in allen vier Zeitschriften beobachten ließ, war unerwartet und ist schwerer zu erklären. Die Unterschiede in der Stichprobengröße waren im Allgemeinen so gering, dass eine absichtliche Manipulation unwahrscheinlich ist; wahrscheinlicher ist, dass Patienten, die nach Studienbeginn ausgeschieden sind, im Studienbericht nicht erwähnt wurden. Studienabbrüche sind häufig durch Nebenwirkungen bedingt, welche mit größerer Wahrscheinlichkeit in der experimentellen Gruppe auftreten. Unter den von Lavori et al. [4] besprochenen 47 Studien waren nur 15 randomisierte Studien ohne Blockbildung, bei denen wir auf eine ähnliche Asymmetrie stießen. In zwei Studien waren die Stichprobenumfänge in beiden Gruppen ähnlich groß, und acht der übrigen 13 Studien (62%) wiesen mehr Patienten in der Kontrollgruppe auf.

Ausgangscharakteristika

Da die randomisierte Behandlungszuteilung zu zufallsbedingten Schwankungen zwischen den Gruppen führen kann, sollte der Grad der erreichten Ähnlichkeit nachgewiesen werden. Die Anzahl der angegebenen Vergleiche von Ausgangswerten schwankte beträchtlich: Zwei Drittel der Studien machten Angaben zu mehr als fünf Variablen, und sechs Studien enthielten dazu keinerlei Informationen. Bei kontinuierlichen Daten sind neben Informationen über deren Variabilität (z. B. Standardabweichung, Bereich, ausgewählte Perzentile oder manchmal auch alle Daten) auch Angaben zu Mittelwert oder Median relevant. 33 Studien (48%) enthielten jedoch keine oder nur unzureichende Angaben zu diesen Streuungsmaßen. Der Standardfehler ist kein deskriptives Maß, sondern bezeichnet vielmehr die Unsicherheit eines Schätzers wie des Mittelwertes [17]. So gesehen sollten weder der Standardfehler noch das damit eng verwandte Konfidenzintervall verwendet werden, wenn Angaben zu den Ausgangswerten gemacht werden. Häufig wurden Hypothesentests angewendet, doch nicht immer wurden auch die Methoden spezifiziert. Zur Bewertung von Ähnlichkeit sind Hypothesentests jedenfalls nicht geeignet; eine solche Bewertung sollte sich auf die prognostische Stärke der Variablen und die Größe des Ungleichgewichts stützen [4, 6]. Bei angemessener Randomisierung ist die Nullhypothese, dass die beiden Gruppen derselben Grundgesamtheit entstammen, per definitionem wahr; demnach würden wir erwarten, dass 5% solcher Vergleiche auf einem Niveau von 5% signifikant sind. Diese Tests bewerten also indirekt, ob die Randomisierung angemessen war und nicht, ob die beiden Gruppen ähnliche Merkmale aufwiesen. Wenn wir all diese Tests zusammen nehmen, waren 4% von 600 Vergleichen auf einem Niveau von 5% signifikant. Damit konnte nicht nachgewiesen werden, dass dafür andere, nicht zufallsbedingte Gründe verantwortlich waren. Wir haben nach Hinweisen gesucht, dass die Autoren in Betracht gezogen hatten, ob Unterschiede zwischen den Gruppen sich auf den Vergleich zwischen den Behandlungen auswirken könnten. Wenn die Gruppen sich im Hinblick auf die prognostischen Variablen ähnlich sind, kann die Analyse einfach sein. Wenn jedoch Unterschiede bestehen, die potenziell wichtig sein könnten, sollte die Auswertung der Ergebnisse z. B. durch ein Regressionsmodell oder die Untersuchung der Veränderungen gegenüber den Ausgangswerten entsprechend modifiziert werden. Insgesamt wurde dieses Thema in nur etwa 60% der Studien angemessen behandelt.

Empfehlungen

Wenn ein bestimmter Punkt in einer Veröffentlichung nicht angesprochen wird, ist es in der Regel nicht möglich zu entscheiden, ob er bei der Datenauswertung gar nicht berücksichtigt oder im Studienbericht nur nicht erwähnt wurde. Liberati et al. [18] untersuchten die in 63 randomisierten Studien über die Primärversorgung des Mammakarzinoms veröffentlichten Daten und nahmen dazu auch telefonischen Kontakt zu den Studienleitern auf, um verschiedene Fragen zu klären. Vor dem Hintergrund dieser Zusatzinformationen stieg der Anteil der Studien, bei denen man von einer angemessen verblindeten Randomisierung ausgehen konnte, von 25% auf 43% und der Anteil derer, die zur Festlegung der Stichprobengröße die Studienpower berechnet hatten, von 32% auf 52%. Diese Ergebnisse lassen darauf schließen, dass sich zwar eine ganze Reihe von Studien in ihren Studienberichten unter Wert verkauft, dass aber in der Mehrzahl der Studien nicht etwa eine lü-

ckenhafte Darstellung für die fehlenden Angaben verantwortlich ist, sondern die Tatsache, dass die Studienautoren die dafür nötigen Maßnahmen gar nicht durchgeführt haben. In unserer Untersuchung wurden in allen vier Fachzeitschriften wichtige Informationen über die Studienmethoden gewöhnlich weggelassen.

Es liegen zahlreiche negative Berichte über die Qualität der Datenpräsentation in medizinischen Fachzeitschriften, insbesondere im Zusammenhang mit klinischen Studien, vor [1, 3, 13, 14, 16, 18–20]. Unsere Studie hat ergeben, dass sich dies in einigen Punkten auch 1987 noch nicht geändert hatte. Ein Bericht über eine randomisierte klinische Studie sollte die folgenden statistischen Angaben enthalten: (A) Beschreibung des Studiendesigns (u. a. Art der Randomisierung); (B) Nachweis der randomisierten Behandlungszuteilung (die zur Generierung der Zufallszahlen angewandte Methode); (C) Durchführung der Behandlungszuteilung, u. a. Angaben dazu, ob sie verblindet oder unverblindet erfolgte; (D) Bestimmung der Stichprobengröße und (E) Vergleich der Ausgangswerte und angemessene Handhabung etwaiger Unterschiede. Wichtig ist ferner auch, ob Patienten, Behandler und Bewerter verblindet waren. Der Begriff „doppelblind" bedarf der näheren Ausführung. Selbstverständlich sollten in allen Veröffentlichungen die statistischen Analysemethoden adäquat beschrieben und die Ergebnisse nachvollziehbar ausgewertet werden. Autoren sollte hierfür eine Art Prüfliste zur Verfügung gestellt werden. In vorhandenen Checklisten [9, 11, 21, 22] sind die Punkte „Behandlungszuteilung" und „Vergleich der Ausgangscharakteristika" nicht so umfassend abgedeckt, wie wir uns das vorstellen. Und selbst wenn die Autoren eine solche Checkliste zu ihrer Verfügung haben, gibt es keine Garantie dafür, dass sie auch alle darin aufgeführten Punkte berücksichtigen. Dieselbe Liste könnte auch bei der redaktionellen Bearbeitung eingesetzt werden, was jedoch zeitraubend und ineffizient ist. Besser wäre es, wenn Autoren eine Checkliste ausfüllen und für jeden Punkt Seite und Absatz angeben müssten, in denen die fraglichen Informationen zu finden sind. Ein solches Vorgehen würde eine bessere Berichtskultur fördern, die redaktionelle Bewertung erleichtern und auf diese Weise den Qualitätsstandard veröffentlichter klinischer Studien anheben.

Wir bedanken uns bei Iain Chalmers, Michael Hughes und Tony Johnson für ihre hilfreichen Vorschläge.

Literatur

[1] Altman DG. Statistics in medical journals. Stat Med 1982;1:59–71.

[2] Tyson JE, Furzan JA, Reisch JS, Mize SG. An evaluation of the quality of therapeutic studies in perinatal medicine. J Pediatr 1983;102:10–3.

[3] DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. N Engl J Med 1982;306: 1332–7.

[4] Lavori PW, Louis TA, Bailar JC, Polansky M. Designs for experiments – parallel comparisons of treatment. N Engl J Med 1983;309:1291–9.

[5] Rothman K. Epidemiologic methods in clinical trials. Cancer 1977;39:1771–5.

[6] Altman DG. Comparability of randomised groups. Statistician 1985;34: 125–36.

[7] Hughes WT, Rivera GK, Schell MJ, Thornton D, Lott L. Successful intermittent chemoprophylaxis for Pneumocystis carinii pneumonitis. N Engl J Med 1987;316:1627–32.

[8] Pocock SJ. Clinical trials: a practical approach. Chichester: John Wiley, 1983: 66–99.

[9] Chalmers TC, Smith H, Blackburn B, et al. A method for assessing the quality of a randomized control trial. Controlled Clin Trials 1981;2:31–49.

[10] Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statisical guidelines for contributors to medical journals. In: Gardner MJ, Altman DG eds. Statistics with confidence. London: British Medical Journal, 1989: 83–100.

[11] Simon R, Wittes RE. Methodological guidelines for reports of clinical trials. Cancer Treat Rep 1985;69:1–3.

[12] Zelen M. Guidelines for publishing papers on cancer clinical trials: responsibilities of editors and authors. Prog Clin Biol Res 1983;132E:57–68.

[13] Mosteller F, Gilbert JP, McPeek B. Reporting standards and research strategies for controlled trials: agenda for the editor. Controlled Clin Trials 1980;1: 37–58.

[14] Evans M, Pollock AV. Trials on trial: a review of trials of antibiotic prophylaxis. Arch Surg 1984;119:109–13.

[15] Keirse MJNC. Amniotomy or oxytocin for induction of labor: re-analysis of a randomized controlled trial. Acta Obstet Gynecol Scand 1988;67:731–5.

[16] Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. N Engl J Med 1987;317:426–32.

[17] Altman DG, Gardner MJ. Presentation of variability. Lancet 1986; ii: 639.

[18] Liberati A, Himel HN, Chalmers TC. A quality assessment of randomised controlled trials of primary treatment of breast cancer. J Clin Oncol 1986;4: 942–51.

[19] Meinert CL, Tonascia S, Higgins K. Content of reports on clinical trials: a critical review. Controlled Clin Trials 1984; 5:328–47.

[20] Götzsche P. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. Controlled Clin Trials 1989; 1031–56.

[21] Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content of medical studies. In: Gardner MJ, Altman DG, eds. Statistics with confidence. London: British Medical Journal, 1989:101–8.

[22] Grant A. Reporting controlled trials. Br J Obstet Gynaecol 1989;96:397–400.

Korrespondenzadresse:
D.G. Altman
Cancer Research UK/NHS Centre for Statistics in Medicine, Wolfson College, Oxford OX2 6UD, Großbritannien
doug.altman@cancer.org.uk

Anmerkung der Redaktion:
Die Übersetzung dieses Artikels erfolgte durch Frau Karin Beifuss (Stuttgart), die fachliche Bearbeitung übernahm Frau Gerta Rücker (IMBI – Institut für Medizinische Biometrie und Statistik Universitätsklinikum Freiburg). Beiden sei an dieser Stelle sehr herzlich gedankt.