5 Tricks bei der Auswertung und Präsentation von Studiendaten

Sie erhalten von einem Phytopharmaka-Hersteller einen Werbeprospekt, in dem die Ergebnisse einer neuen Studie zur Demenzprävention zusammengefasst werden. Dort lesen Sie: „Zwar wurde der primäre Endpunkt nicht erreicht. Eine hochsignifikante Risikoreduktion fand sich jedoch bei Männern, die das Mittel über einen Zeitraum von vier Jahren eingenommen hatten."

Systematische Verzerrungen können nicht nur beim Design und der Durchführung einer Studie auftreten, sondern auch bei der Auswertung und Darstellung der Daten. Denn Studienautoren und Hersteller sind bei der Berechnung und Präsentation der Therapieeffekte nicht in allen Fällen auch tatsächlich objektiv.

Beim kritischen Lesen von klinischen Studien sollte man deshalb auf einige Aspekte achten, die zu fehlerhaften Schlussfolgerungen führen können. Das betrifft etwa die Frage, wie die Autoren mit den Daten von Studienabbrechern bei der Auswertung umgehen (▸Kap. 5.1), wie sie statistische Fallstricke wie Probleme des multiplen Testens handhaben (▸Kap. 5.2) und ob sie bestimmte Aspekte bei Nicht-Unterlegenheitsstudien beachten (▸Kap. 5.3).

Auf problematische Darstellungen von Studienergebnissen trifft man besonders häufig in der Arzneimittelwerbung: Nicht immer gibt der Werbeprospekt tatsächlich auch die relevanten Daten wieder. In manchen Fällen sind die Ergebnisse auch so dargestellt, dass die Wahrnehmung des Lesers in eine bestimmte Richtung gelenkt wird – bei objektiver Betrachtung lässt sich die Werbeaussage mit den zitierten Daten aber gar nicht sicher belegen (▸Kap. 5.4).

In den letzten Jahren wird auch vermehrt das Problem thematisiert, dass unliebsame Studienergebnisse vielfach nicht veröffentlicht werden. Hintergründe und aktuelle Entwicklungen zu dieser Thematik finden sich in einem kleinen Exkurs (▸Kap. 5.5).

5.1 Umgang mit Studienabbrechern

Nicht immer hält sich das Leben an den Plan – das gilt auch für klinische Studien. So wird es auch bei guter Studienplanung und sorgfältiger Betreuung Patienten geben, die die Therapie nicht wie verordnet durchführen. Manche erscheinen vielleicht nicht zu einzelnen vorgesehenen Untersuchungen oder brechen die Studie sogar vorzeitig ab.

Die Gründe dafür sind vielfältig: So werden den Patienten vielleicht die häufigen Untersuchungen im Rahmen der Studie zu viel und sie widerrufen ihre Einwilligung. Eventuell verziehen sie in eine andere Stadt oder versterben aus Gründen, die nichts mit der Studie zu tun haben (etwa bei einem Verkehrsunfall). Jedoch gibt es vielleicht auch Patienten, die nicht mehr an der Studie teilnehmen wollen, weil sie die Nebenwirkungen

nicht aushalten oder von der ausbleibenden Linderung ihrer Beschwerden enttäuscht sind. Bei manchen Patienten entscheidet vielleicht sogar der Arzt, dass ihnen eine weitere Behandlung im Rahmen der Studie nicht zumutbar ist. Wie geht man jetzt mit den Daten dieser Patienten (Drop-outs) bei der Auswertung der Studie um?

Ein natürlicher Impuls wäre es, nur die Daten von denjenigen Patienten auszuwerten, die gemäß dem Studienprotokoll und bis zum Ende der Studie behandelt wurden. Dieses Vorgehen heißt „per-protocol"(PP)-Analyse. Weil der Studienabbruch meist im Zusammenhang mit der Therapie steht, sind die ausscheidenden Patienten in der Regel nicht gleichmäßig auf Behandlungs- und Kontrollgruppe verteilt. Bei einer per-protocol-Analyse würde dann das Ergebnis der Studie systematisch verzerrt, in vielen Fällen zugunsten der Behandlungsgruppe.

In manchen Fällen wechseln Patienten auch während der Studie die Gruppen. Wertet man die Patienten in der Gruppe aus, in der sie tatsächlich behandelt wurden, spricht man von einer „as-treated"-Analyse. Das hat aber den Nachteil, dass die Randomisierung nicht aufrechterhalten wird und sich so ebenfalls systematische Verzerrungen einschleichen können.

Aus diesem Grund ist es in den meisten Fällen sinnvoll und notwendig, die Auswertung in den Gruppen vorzunehmen, in denen die Patienten ursprünglich randomisiert waren. Dieses Vorgehen heißt „intention-to-treat"(ITT)-Analyse. Auch Patienten, die zwar randomisiert wurden, aber keine Behandlung erhalten haben, sollten in die Analyse mit einbezogen werden.

In der Studienpublikation sollte genau beschrieben werden, wie die Daten der Patienten im Hinblick auf die Gruppenzugehörigkeit ausgewertet werden. Denn Begriffe wie „modifizierte intention-to-treat-Analyse" sind nicht standardisiert und werden für verschiedene Vorgehensweisen genutzt, bei denen es sich beispielsweise auch im Wesentlichen um eine per-protocol-Analyse handeln kann. Hilfreich ist deshalb eine schematische Darstellung nach dem CONSORT-Statement, auf der der Patientenfluss im Verlauf der Studie nachgezeichnet wird. So lässt sich auch leicht nachvollziehen, an welcher Stelle Patienten aus der Studie ausgeschieden sind (Abb. 5.1).

Die Forderung nach einer intention-to-treat-Analyse gilt in erster Linie für die Auswertung der Wirksamkeit in einer Überlegenheitsstudie, wenn also nachgewiesen werden soll, dass ein neues Medikament besser wirkt als die bisherige Standardtherapie. Bei Äquivalenz- oder Nicht-Unterlegenheitsstudien gelten andere Regeln (Kap. 5.3). Um Nebenwirkungen zu analysieren, wird die as-treated-Analyse bevorzugt.

Allgemein akzeptiert wird als Ausnahme von der Regel einer intention-to-treat-Analyse, wenn sich nach der Randomisierung herausstellt, dass der Patient trotz aller Sorgfalt doch nicht die Einschlusskriterien für die Studie erfüllt. Um solche Patienten dann nachträglich aus der Studie auszuschließen, sind jedoch eine Reihe von statistischen Vorsichtsmaßnahmen nötig, um die Ergebnisse nicht zu verzerren.

Die intention-to-treat-Analyse ist ein konservativer Ansatz, der den Effekt der Behandlung eher unterschätzt. Allerdings halten sich in der Praxis auch nicht alle Patienten an die Therapievorschrift, so dass die intention-to-treat-Analyse zum Teil auch die Alltagsbedingungen abbildet. Per-protocol- und as-treated-Analysen führen dagegen eher zu einer Überschätzung des Therapieeffekts. Interessant ist es in vielen Fällen auch, die Unterschiede der Effektschätzer anzuschauen, je nachdem welche Art der Datenanalyse verwendet wird. Aus diesem Grund finden sich in Studienpublikationen auch häufig Sensitivitätsanalysen, bei denen die Ergebnisse der Analysenmethoden miteinander verglichen

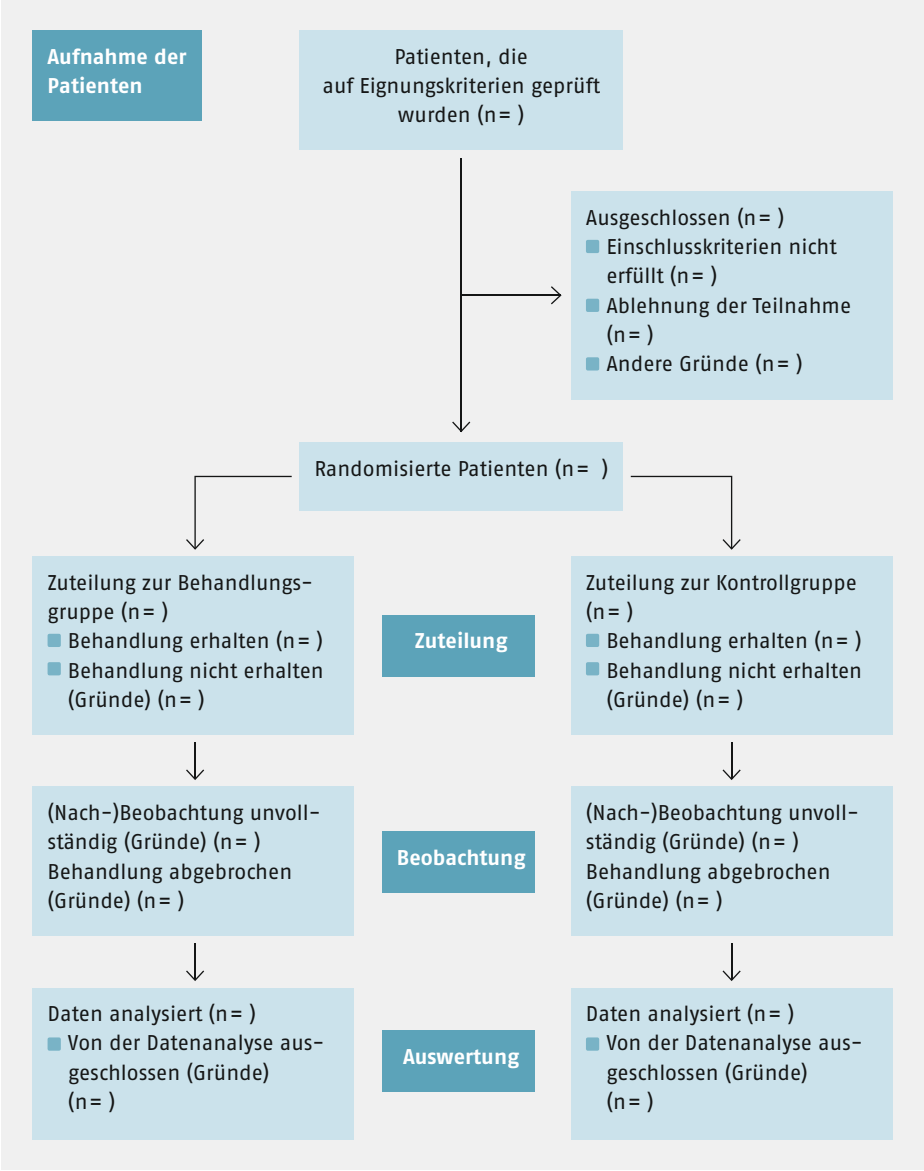

Abb. 5.1 Beispiel für ein Flussdiagramm entsprechend dem CONSORT-Statement

werden. Im Idealfall gibt es zwischen den Ergebnissen der beiden Analysenmethoden keinen wesentlichen Unterschied.

In der Publikation sollten die Autoren außerdem angeben, wie sie mit den fehlenden Daten der Studienabbrecher in der Auswertung umgegangen sind. Dafür sind verschiedene statistische Herangehensweisen gebräuchlich. Dazu gehört etwa „last observation carried forward", also das Einsetzen des letzten beobachteten Wertes im weiteren Verlauf der Studie. Dadurch kann allerdings die Wirksamkeit des Mittels in der Behandlungsgruppe überschätzt werden. Eine mögliche konservative Alternative ist es, bei binären

Zielgrößen die Daten von ausgeschiedenen Studienteilnehmern als Therapieversagen zu werten. Das unterschätzt allerdings die Wirksamkeit der Behandlung. Gerne werden deshalb auch in Studienpublikationen „best case"- und „worst case"-Szenarien entwickelt, die die Wirksamkeit der Therapie auf der Basis von möglichen extremen Resultaten abschätzen.

Bei einer hohen Anzahl von Studienabbrechern können solche Abschätzungen jedoch sehr unzuverlässig werden. Als Faustregel gilt, dass bei Beobachtungsverlusten von mehr als 20 % die Studienergebnisse nicht mehr als valide angesehen werden können. In Einzelfällen können jedoch bessere Nachbeobachtungsquoten aufgrund des Settings nicht erreicht werden oder bei sehr seltenen Ereignissen auch schon geringere Verluste problematisch sein. Die Autoren sollten die genannten Aspekte von fehlenden Daten in der Diskussion der Ergebnisse entsprechend berücksichtigen. Bei der Lektüre von Studienpublikationen sollte man besonders aufmerksam werden, wenn sich die Verlustraten in Behandlungs- und Kontrollgruppe deutlich unterscheiden.

5.2 Probleme des multiplen Testens

Wenn man für den Aufwand einer klinischen Studie eine maximale Ausbeute erreichen will, sollte man so viele Endpunkte wie möglich untersuchen. Außerdem sollte man in den erhobenen Daten danach graben, ob es nicht doch einzelne Patientengruppen gibt, bei denen das Testpräparat deutlich besser wirkt. Für jedes Teilergebnis kann man dann paarweise statistische Tests durchführen und signifikante Ergebnisse mit Fettdruck in der Publikation markieren. Und dann ist es auch wichtig, die Ergebnisse zu möglichst vielen

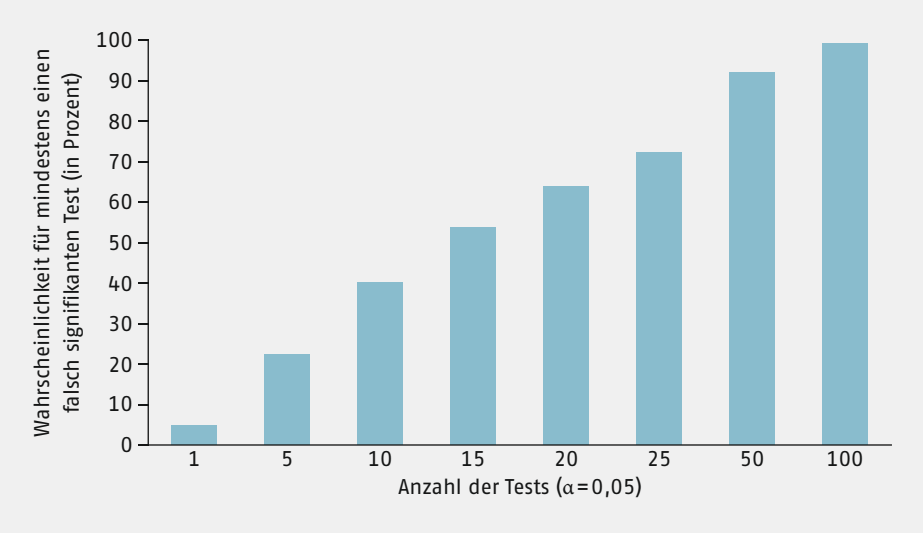

Abb. 5.2 Wie hoch die Wahrscheinlichkeit dafür ist, bei einer bestimmten Anzahl statistischer Tests fälschlicherweise mindestens ein signifikantes Ergebnis zu erhalten, lässt sich berechnen. Bei einem Signifikanzniveau von 5 % (α = 0,05) liegt bereits bei 15 statistischen Tests die Wahrscheinlichkeit bei mehr als 50 %.

Zeitpunkten zu analysieren und die Studie so schnell wie möglich abzubrechen, sobald das neue Arzneimittel statistisch signifikant bessere Ergebnisse zeigt.

Diese Parodie ist ein Albtraum für Statistiker. Denn sie ignoriert grundlegend das Problem des multiplen Testens, das bei einer Vielzahl von Studien auftreten kann. Das hängt zum einen damit zusammen, dass die mit großen Investitionen erhobenen Daten möglichst gut ausgenutzt werden sollen. Zum anderen lassen sich Ergebnisse mit dem Zusatz „signifikant" deutlich besser vermarkten – unabhängig davon, wie groß tatsächlich der Effekt ist (zum Begriff „signifikant" vergleiche auch ▸Kap. 4.5.2). In der Praxis führt das dazu, dass in Studien häufig eine Vielzahl von statistischen Tests an den gleichen Daten durchgeführt werden (post-hoc-Tests). In vielen Fällen führt das aber zu einem erhöhten Risiko für ein zufällig signifikantes Ergebnis, das in Wirklichkeit nicht existiert. Anders ausgedrückt: Durch unsachgemäßes multiples Testen liegt die Irrtumswahrscheinlichkeit für das Gesamtergebnis nicht mehr bei den häufig angestrebten 5 %, sondern deutlich darüber (◉ Abb. 5.2).

Das multiple Testen wirft also eine Reihe von Problemen auf. Wie man allerdings richtig damit umgeht, ist unter Statistikern nicht ganz unumstritten. So gibt es etwa verschiedene statistische Methoden (etwa nach Bonferroni oder Holm), bei denen die Irrtumswahrscheinlichkeit für die Einzelhypothesen so adjustiert wird, dass für das Gesamtergebnis dann wieder die richtige „globale" Irrtumswahrscheinlichkeit gilt. Das hat verschiedene Vor-, aber auch Nachteile. Für die verschiedenen Situationen, in denen das Problem des multiplen Testens auftreten kann, werden deshalb unterschiedliche Strategien vorgeschlagen. Grundsätzlich gilt aber: Die Anzahl der statistischen Tests sollte so gering wie möglich gehalten werden.

> *„Wenn man die Daten lange genug foltert, werden sie etwas gestehen"*
> zugeschrieben dem Ökonomen Ronald Coase (1910–2013)

5.2.1 Endpunkte

Werden in einer Studie mehrere Endpunkte untersucht, sollten die Autoren der Studien einen als den primären Endpunkt auswählen. Im Idealfall ist das diejenige Zielgröße, die für die Patienten am relevantesten ist. Für diesen primären Endpunkt erfolgt die Fallzahlberechnung und die Ergebnisse dieses Endpunktes werden mit Hilfe eines statistischen Tests auf Signifikanz untersucht (konfirmatorischer Test). Das Ergebnis dieses Tests kann dann auch zu recht als statistisch signifikant bezeichnet werden, wenn der p-Wert unter 0,05 liegt. Die Auswertung der sekundären Endpunkte erfolgt dann rein explorativ, niedrige p-Werte sind dann also statistisch nicht aussagekräftig und müssen in weiteren Studien als primärer Endpunkt untersucht werden, wenn ein konfirmatorischer statistischer Test durchgeführt werden soll.

Bei der Lektüre von Studien sollte man als Leser darauf achten, ob sich die behaupteten signifikanten Effekte tatsächlich auch für den primären Endpunkt finden. Misstrauisch sollte man werden, wenn sich in einer Studie kein Unterschied bei dem primären Endpunkt feststellen lässt, die Autoren sich in der Diskussion aber auf Unterschiede bei den sekundären Endpunkten konzentrieren.

5.2.2 Subgruppenanalysen

Ähnliches gilt auch für Subgruppenanalysen, also die Auswertung der Studiendaten für bestimmte Untergruppen der eingeschlossenen Patienten. Diese sind besonders dann problematisch, wenn die Untergruppen erst nach der Erhebung der Daten gebildet wurden und nur jeweils eine kleine Stichprobe von Patienten mit wenigen Ereignissen enthalten. Auch wenn sehr ungleiche Gruppengrößen entstehen, können Probleme auftreten. Dann können sowohl Zufallseffekte als auch systematische Verzerrungen durch Verletzung der Randomisierung (Selektionsbias) die gefundenen Unterschiede zwischen den Subgruppen erklären (▸Kap. 3.2).

Deshalb gibt es einige Anforderungen an Subgruppenanalysen, wenn man zu zulässigen Aussagen kommen will: So fordern Experten, dass Subgruppenanalysen bereits vor Beginn der Studie spezifiziert sein müssen. Sinnvolle Subgruppen beziehen sich auf Parameter, die bereits vor Beginn der Studie vorhanden waren, etwa Altersgruppen oder bestimmte Risikokonstellationen. Im Idealfall lassen sich solche Subgruppen dann auch bei der Fallzahlplanung berücksichtigen, etwa im Rahmen einer stratifizierten Randomisierung. Gibt es Hinweise darauf, dass beispielsweise die Wirksamkeit eines Medikaments bei Frauen anders sein könnte als bei Männern oder sich mit zunehmendem Alter verändert, sollte die Studie von vorneherein entsprechend angelegt sein. Wichtig ist auch ein angemessenes statistisches Auswerteverfahren, etwa ein Interaktionstest statt einer Vielzahl von paarweisen Vergleichen.

Wenn man sich nicht an diese Regeln hält, können abstruse Resultate entstehen. In einem berühmten Beispiel haben Autoren das einmal im Extrem durchexerziert, um die Gefahren von Subgruppenanalysen zu verdeutlichen: In einer Untersuchung zur Sekundärprophylaxe eines Herzinfarkts wurden anhand willkürlicher Zuordnungen Subgruppen gebildet und entsprechende statistische Hypothesentests durchgeführt. Das Ergebnis: Patienten, die unter den Sternzeichen Zwillinge oder Waage geboren waren, hatten keinen Nutzen durch die Gabe von Acetylsalicylsäure, während Patienten mit anderen Sternzeichen signifikant davon profitierten. Leider lässt sich in vielen anderen Studien nicht so leicht wie in diesem Beispiel durchschauen, dass die Ergebnisse der Subgruppenanalyse sehr wahrscheinlich Zufallsbefunde sind.

Eine biologische Erklärung, warum das Mittel bei einer bestimmten Subgruppe besser wirkt als bei einer anderen, kann die Glaubwürdigkeit von Subgruppenanalysen erhöhen. Allerdings ist auch dann Vorsicht geboten. Wenn ein Subgruppeneffekt konsistent in mehreren Studien nachgewiesen wurde, kann das ein Anzeichen dafür sein, dass ein solcher Effekt tatsächlich existiert. Deshalb halten Experten Replikation in verschiedenen Studien auch für einen aussagekräftigeren Test auf Subgruppeneffekte als statistische Signifikanztests. Im Idealfall sollten Subgruppeneffekte in RCT bestätigt werden, die für die Fragestellung entsprechend angelegt sind.

5.2.3 Zwischenauswertungen und vorzeitiger Studienabbruch

In manchen Fällen werden klinische Studien vorzeitig abgebrochen. Sie enden also, bevor die ursprünglich geplante Nachbeobachtungszeit verstrichen ist und/oder bevor die geplante Anzahl an Teilnehmern in die Studie aufgenommen wurde. Die Gründe dafür können vielfältig sein:

So kann eine Zwischenauswertung der Ergebnisse zu der vorzeitigen Schlussfolgerung führen, dass eine der beiden untersuchten Therapien deutlich effektiver oder besser verträglich sei. Dann könnten ethische Gesichtspunkte dafür sprechen, den Patienten in der

anderen Gruppe die „bessere" Therapie nicht vorzuenthalten. Begünstigt werden solche Entscheidungen auch dadurch, dass sich solche positiven Ergebnisse besser publizieren und die entsprechenden Arzneimittel besser verkaufen lassen. Kürzere Studien sind auch deutlich kostengünstiger als solche mit längerer Laufzeit.

Allerdings birgt ein vorzeitiger Studienabbruch auch zahlreiche Schwierigkeiten, die teilweise mit dem Problem des multiplen Testens zusammenhängen. So wird mehrfach im Verlauf der Studie getestet, ob signifikante Unterschiede zwischen den Behandlungsgruppen im Hinblick auf erwünschte und unerwünschte Wirkungen bestehen. Bei der Datenanalyse müssen dann aber auch die mehrfachen Tests im Hinblick auf das Signifikanzniveau entsprechend berücksichtigt werden.

Hinzu kommt, dass besonders zu Beginn einer Studie Therapieeffekte zufallsbedingt stark schwanken können. Wird die Studie zu einem sehr frühen Zeitpunkt gestoppt und beruht die Entscheidung dazu auf relativ wenigen Ereignissen, besteht ein hohes Risiko, dass die Therapieeffekte deutlich überschätzt werden.

Empfohlen wird deshalb, dass in klinischen Studien nur ein unabhängiges Gremium (Data and Safety Monitoring Board, auch DSMB abgekürzt) während der Laufzeit der Studie Zugang zu den unverblindeten Studiendaten hat. Das DSMB sollte die Daten nach vorher festgelegten statistischen (Stopp-)Regeln bewerten, die die beschriebenen Probleme berücksichtigen.

Daneben kann sich natürlich in einer klinischen Studie durch eine vorzeitige Beendigung auch zusätzlich zu den statistischen Problemen noch die Schwierigkeit ergeben, dass man den Verlauf nicht sicher vorhersagen kann und damit ein hohes Risiko für falsche Schlussfolgerungen entsteht.

> *„Nach fünf von sechs Versuchen mit dem Revolver waren sich die Wissenschaftler einig, dass Russisch Roulette vollkommen ungefährlich ist."*

Schließlich ist auch zu bedenken, dass besonders bei Therapien, die langfristig angewendet werden (etwa bei Diabetes oder Hypertonie) ein vorzeitiger Studienabbruch auch dazu führen kann, dass aus der Studie keine Schlussfolgerungen über erwünschte oder unerwünschte Effekte nach längerer Behandlungsdauer gezogen werden können.

5.2.4 Hilfe für die Bewertung von Problemen mit dem multiplen Testen

In der Vergangenheit gab es nicht selten Fälle, dass die Untersucher einer Studie mit den unterschiedlichen Endpunkten eine ganze Reihe von statistischen Tests durchgeführt haben und nur diejenigen publiziert, bei denen signifikante Unterschiede gefunden wurden. Gleiches fand sich auch häufig bei Subgruppenanalysen oder Studien, die zu einem günstigen Zeitpunkt abgebrochen wurden. Das CONSORT-Statement, der Standard für Publikationen von randomisierten kontrollierten Studien, fordert deshalb, dass die prospektive Planung aller entsprechenden Vorgehensweisen im Methodenteil beschrieben wird.

Ob es bei der Auswertung der Studiendaten nachträglich Änderungen gegeben hat, lässt sich letzten Endes aber nur durch den Vergleich der Publikation mit dem Studienprotokoll im entsprechenden Studienregister nachweisen. Diese Abklärung ist im Detail

für den einzelnen Leser aber meist nicht zu leisten und sollte eigentlich durch die Gutachter der Zeitschrift erfolgen, in der die Studie publiziert wurde.

Der Leser kann deshalb in der Regel nur prüfen, ob sich entsprechend dem CONSORT-Statement bestimmte Angaben im Methodenteil finden und in der Diskussion entsprechend gewichtet werden. Dazu gehört etwa die Angabe, welcher Endpunkt als der primäre Endpunkt gilt, für den die Fallzahlplanung durchgeführt wird. Auch sollten im Methodenteil die Subgruppenanalysen aufgeführt werden, die vorab geplant waren. Details zur vorgesehenen Studiendauer und der vorgesehenen Teilnehmerzahl sind ebenfalls notwendig – eine kleinere als die vorgesehene Probandenzahl kann ein Indiz für einen vorzeitigen Studienabbruch sein, auch wenn es in der Studie nicht explizit erwähnt wird. Die Autoren sollten auch erläutern, nach welchen statistischen Verfahren Regeln für einen eventuell notwendigen vorzeitigen Abbruch der Studie aufgestellt wurden. Probleme des multiplen Testens von statistischen Hypothesen sollten die Autoren sowohl im Methodenteil als auch in der Darstellung der Resultate und der zugehörigen Diskussion entsprechend berücksichtigen.

5.3 Nicht-Unterlegenheitsstudien

Die Studien, die bisher betrachtet wurden, beschäftigten sich mit der Frage, ob ein neues Arzneimittel besser wirkt als die bisherige Standardtherapie. Das entsprechende Design bezeichnet man auch als Überlegenheitsstudie. Wie bereits in ▸Kap. 4.5.1 erläutert, werden für einen statistischen Test dann eine Nullhypothese sowie eine Alternativhypothese aufgestellt. Bei Überlegenheitsstudien lautet die entsprechende Nullhypothese bei zweiseitiger Fragestellung: Die beiden Mittel wirken gleich gut, es gibt keinen Unterschied. Die Alternativhypothese besagt: Es gibt einen Unterschied zwischen den beiden Mitteln. Diese Konstruktion wird notwendig, weil man im statistischen Sinn eine Hypothese nicht beweisen, sondern nur widerlegen beziehungsweise nicht widerlegen kann.

Nun gibt es aber auch wissenschaftliche Fragestellungen, bei denen man nicht die Überlegenheit eines neuen Mittels nachweisen will. Vielmehr will man zeigen, dass ein anderes Mittel oder eine andere Behandlung gleich gut wie die Standardtherapie oder ihr zumindest nicht unterlegen ist. Dafür kann es verschiedene Gründe geben, beispielsweise wenn zu vermuten ist, dass das neue Mittel ungefähr gleich gut wirkt wie der ältere Wirkstoff, aber weniger Nebenwirkungen hat, für Patienten mit Kontraindikationen gegen den älteren Wirkstoff geeignet ist, leichter anzuwenden oder deutlich preisgünstiger ist. Für viele Indikationen gibt es außerdem inzwischen eine gut wirksame Standardtherapie, so dass es sehr schwer wäre, eine Überlegenheit nachzuweisen.

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

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

und Auswertung, die auch in einer separaten Erweiterung des CONSORT-Statements festgehalten sind.

Das beginnt bereits bei der Formulierung der Hypothesen, die bei Nicht-Unterlegenheitsstudien umgekehrt wie bei einer Überlegenheitsstudie aufgestellt werden. Die Nullhypothese heißt also: „Das alte Arzneimittel wirkt besser als das neue", während die Alternativhypothese besagt: „Das neue Arzneimittel wirkt nicht schlechter als das alte". Das ist deshalb wichtig, weil der „Nachweis von Nicht-Unterlegenheit des neuen Arzneimittels" im statistischen Sinn nicht das gleiche ist wie „Die Überlegenheit des alten Arzneimittels kann nicht nachgewiesen werden". Denn „kein Nachweis der Überlegenheit" würde sich beispielsweise auch ergeben, wenn die Ergebnisse zwischen den Teilnehmern sehr stark streuen.

Natürlich muss man auch definieren, was eigentlich genau „Unterschied" heißt. Denn in der Regel wird man bei der Studie nie exakt gleiche Ergebnisse für die beiden Arzneimittel erhalten, sondern man muss auch immer auch mit zufälligen Einflüssen rechnen. Bei Nicht-Unterlegenheitsstudien werden deshalb Nicht-Unterlegenheitsgrenzen festgelegt, die den Einfluss des Zufalls kontrollieren. Diese Festlegung erfolgt vor Studienbeginn, also bevor die Daten erhoben werden.

Bei der Auswertung der Studiendaten gilt das Prinzip: Wenn sich die Differenz zwischen Behandlungs- und Kontrollgruppe vollständig oberhalb der Nicht-Unterlegenheitsgrenze befindet, geht man davon aus, dass die beiden Medikamente die gleiche Wirksamkeit aufweisen. Diese Auswertung erfolgt in der Regel auf der Basis der entsprechenden Konfidenzintervalle. Das CONSORT-Statement empfiehlt, die Äquivalenz- bzw. Nicht-Unterlegenheitsgrenzen und die Konfidenzintervalle in einer Grafik darzustellen, damit der Leser die Ergebnisse leicht nachvollziehen kann (Abb. 5.3).

Für den Nachweis von Nicht-Unterlegenheit ist aber auch eine angemessene Fallzahlplanung wichtig. Eine zu geringe Fallzahl könnte nämlich dazu führen, dass ein in Wirklichkeit bestehender Unterschied aus statistischen Gründen in der Studie nicht erkannt wird. Werden also zu wenig Patienten in die Studie eingeschlossen, steigt das Risiko für einen Fehler 2. Art – man geht also fälschlicherweise davon aus, dass kein Unterschied besteht, obwohl in Wirklichkeit einer vorhanden ist.

Bei einer Nicht-Unterlegenheitsstudie sollte man als Leser auch immer einen wachsamen Blick auf die Vergleichsbehandlung werfen. Denn wenn die Patienten in der Kontrollgruppe etwa nicht mit einer ausreichend hohen Dosis behandelt werden, ist es deutlich leichter, eine Nicht-Unterlegenheit eines neuen Arzneimittels nachzuweisen. Natürlich muss es auch klar sein, dass die Kontrollbehandlung tatsächlich wirksamer als Placebo ist. Denn sonst hieße Äquivalenz im Wesentlichen nur: Wir sind uns sicher, dass das neue Arzneimittel nicht schlechter als eine Placebobehandlung wirkt.

Ein weiterer möglicher Trick besteht in der Auswahl der Patienten: So sollten die Patienten in der Nicht-Unterlegenheitsstudie denen möglichst ähnlich sein, bei denen die Wirksamkeit der Kontrollbehandlung nachgewiesen wurde. Denn untersucht man therapeutische Effekte etwa nur bei Patienten in leichteren Krankheitsstadien, kann man eine Nicht-Unterlegenheit des neuen Mittels leichter nachweisen.

Wichtig zu wissen: Abweichungen vom Studienprotokoll können dazu führen, dass eventuell vorhandene Unterschiede zwischen den untersuchten Medikamenten verwischen. Bei einer Überlegenheitsstudie wird deshalb die Intention-to-treat-Analyse als konservativer Ansatz betrachtet. Bei Nicht-Unterlegenheitsstudie ist es jedoch genau umgekehrt: Hier kann die intention-to-treat-Analyse leicht in die Irre führen. Deshalb

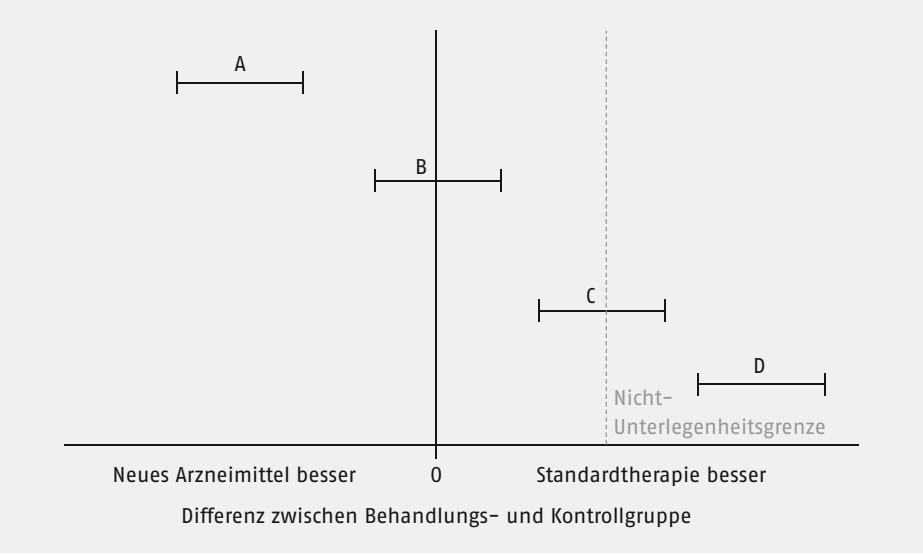

Abb. 5.3 Beispiel für die Beurteilung von Nicht-Unterlegenheit. Gezeigt sind jeweils die Konfidenzintervalle für die Differenz des Therapieeffekts zwischen Behandlungs- und Kontrollgruppe. In den Fällen A und B ist Nicht-Unterlegenheit des Mittels in der Behandlungsgruppe gezeigt, da sich die Konfidenzintervalle vollständig links von der Nicht-Unterlegenheitsgrenze (gestrichelte Linie) befinden (zugunsten der Behandlungsgruppe). In Fall C ist keine eindeutige Beurteilung möglich, da das Konfidenzintervall die Nicht-Unterlegenheitsgrenze einschließt. In Fall D kann nicht auf Nicht-Unterlegenheit des Mittels in der Behandlungsgruppe geschlossen werden, da sich das Konfidenzintervall vollständig diesseits der Nicht-Unterlegenheitsgrenze (zugunsten der Kontrollgruppe) befindet (nach CONSORT).

sollten die Untersucher zusätzlich eine per-protocol-Analyse durchführen, die bei Nicht-Unterlegenheitsstudie den „worst case" abbildet. Im Idealfall sollten die beiden Analysen nicht gravierend voneinander abweichen.

5.4 Fallstricke in der Werbung

Vorsicht ist auch Werbeaussagen zu Arzneimitteln und Nahrungsergänzungsmitteln geboten. Inzwischen gibt es zahlreiche Untersuchungen, die belegen, dass der Leser durch die Werbung leicht zu fehlerhaften Schlussfolgerungen gelangen kann.

So finden sich für Aussagen zur Wirkung des Arzneimittels in vielen Fällen entweder gar keine Belege oder die entsprechenden Untersuchungen liegen nur dem Hersteller vor, wurden aber nicht veröffentlicht (data on file). In solchen Fällen ist es nicht nachzuvollziehen, ob die beworbenen Effekte tatsächlich in Studien nachgewiesen wurden.

Auch wenn als Beleg für die Werbeaussage Studien zitiert werden, lohnt es sich, genauer hinzuschauen. Denn nicht immer sind es tatsächlich randomisierte kontrollierte Studien, sondern manchmal auch Anwendungsbeobachtungen oder Laborversuche, die zitiert werden, ohne es entsprechend kenntlich zu machen. Als erster Anhaltspunkt kann dann eine Literaturrecherche, etwa in PubMed helfen, sich über die Art der Studie zu orientieren (▸Kap. 9.3.2). Aber auch wenn es sich tatsächlich um eine randomisierte

kontrollierte Studie handelt, lohnt ein genauer Blick: Denn nicht immer liegt tatsächlich eine methodisch hochwertige Studie vor, so dass unter Umständen ein erhebliches Verzerrungspotential besteht (▸Kap. 3). Die methodische Qualität lässt sich allerdings in der Regel nicht im meist frei verfügbaren Abstract prüfen, sondern nur im Volltext.

Auch die Details der Werbeaussagen sollte der Leser mit den Ergebnissen vergleichen. So haben Untersuchungen gezeigt, dass nicht in allen Fällen die Werbeaussage tatsächlich mit dem Ergebnis übereinstimmt. Es ist nicht selten, dass nur ein Teil der Ergebnisse, etwa von nachträglichen Subgruppenanalysen zitiert wird, das Gesamtergebnis aber fehlt. Gelegentlich werden die Ergebnisse in der Werbung auch auf Patientengruppen ausgeweitet, die in der eigentlichen Studie gar nicht untersucht wurden. Vorsicht ist auch bei Aussagen zu Nebenwirkungen geboten („gut verträglich"), die manchmal verharmlost oder ganz verschwiegen werden.

Beliebt ist in der Werbung auch die Verwendung von relativen Risikoangaben statt der absoluten Risikoreduktion, weil die Zahlenwerte in der Regel größer ausfallen (▸Kap. 4.2.5). Diese Größe erlaubt aber keinen Aufschluss darüber, wie häufig das Ereignis tatsächlich war. So kann sich eine relative Risikoreduktion in der Herzinfarktprävention von 75 % etwa berechnen, wenn in der Kontrollgruppe bei 20 % der Patienten ein Herzinfarkt aufgetreten ist, in der Behandlungsgruppe aber nur bei 5 % (absolute Risikoreduktion 15 %). Das gleiche Ergebnis in der relativen Risikoreduktion ergibt sich aber auch, wenn die Ereignisrate in der Kontrollgruppe bei 4 % und in der Behandlungsgruppe bei 1 % liegt (absolute Risikoreduktion 3 %).

Gerne werben Hersteller auch mit hochsignifikanten Unterschieden ihres Arzneimittels zur bisherigen Standardtherapie. Hier sollte der Leser sich aber nicht von niedrigen p-Werten blenden lassen, sondern hinterfragen, ob der Unterschied auch tatsächlich klinisch relevant ist (▸Kap. 4.5.5). Auch bei grafischen Darstellungen ist Vorsicht geboten: So lässt sich bei fehlenden Achsenskalierungen oder Nulllinien nicht beurteilen, wie groß tatsächlich ein dargestellter Unterschied ist. Gleiches gilt auch bei abgeschnittenen Säulen.

Zunehmend häufiger wird auch in der Werbung auf Leitlinienempfehlungen verwiesen. Hinterfragen sollte man im Einzelfall allerdings, wie die jeweilige Empfehlung zustande gekommen ist (▸Kap. 8.4) und sich auch informieren, ob die Aussage in der Werbung auch tatsächlich im Detail mit der jeweiligen Leitlinienempfehlung übereinstimmt.

5.5 Exkurs: Grundsätzliche Probleme bei der Publikation von Studienergebnissen

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

Die Nicht-Veröffentlichung kann dabei verschiedene Dimensionen annehmen: So werden einige Studien (besonders solche mit negativen Ergebnissen) gar nicht veröffentlicht, das gilt auch für Studien, die vorzeitig beendet wurden. Auch nachträgliche Veränderungen bei der Wahl der Endpunkte sowie bei Subgruppenanalysen sind keine Seltenheit. Bei anderen Publikationen dagegen fehlen Daten zu einzelnen Endpunkten oder Nebenwirkungen.