Epidemiologie

Multiplizität in randomisierten Studien I: Endpunkte und Behandlungen

Kenneth F. Schulz und David A. Grimes
Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA

aus: *Lancet* 2005;**365**:1591–5
(Originaltitel: „Multiplicity in randomised trials I: endpoints and treatments")

Zusammenfassung

Probleme mit Multiplizität entstehen, wenn in Studien über das Protokoll hinaus weitere Endpunkte (Zielgrößen) untersucht und viele zusätzliche Vergleiche zwischen den Therapiegruppen angestellt werden. Potenziell können sich aus einer einzigen Studie Tausende von Vergleichen ergeben. Wenn nur die signifikanten Vergleiche berichtet werden, dann gilt ein solches Vorgehen, falls es unabsichtlich geschieht, als unwissenschaftlich; erfolgt es jedoch vorsätzlich, handelt es sich schlicht um Betrug. Tatsächlich müssen in einem Studienbericht alle untersuchten Endpunkte und alle miteinander verglichenen Behandlungen dargelegt werden. Um dem Problem des multiplen Vergleichens gerecht zu werden, befürworten manche Statistiker die Durchführung statistischer Adjustierungen. Einfach ausgedrückt testet man „kein Effekt bei allen untersuchten Primärzielgrößen" versus „ein Effekt bei wenigstens einer dieser Zielgrößen". Im Allgemeinen liefern statistische Adjustierungen für Multiplizität grobe Antworten auf eine irrelevante Frage. Adjustiert werden sollte jedoch dann, wenn sich die Argumentation bei der medizinischen Entscheidungsfindung ausschließlich darauf stützt, dass mindestens einer der vorab festgelegten Primärendpunkte signifikant ist. In diesem Fall lässt sich durch Adjustierung eine Auswertung ohne vorher festgelegte Strategie bis zu einem gewissen Grade retten. Leser sollten mit der Möglichkeit rechnen, dass Analysen unterdrückt werden (Underreporting).

Viele analytische Probleme in Studien haben mit Multiplizität zu tun. Normalerweise gehen Studienautoren mit diesen Problemen verantwortungsbewusst um; doch es kommt auch vor, dass Autoren die Auswirkungen dieser Probleme ignorieren oder nicht bedenken. Umgangssprachlich ausgedrückt quälen manche Wissenschaftler ihre Daten so lange, bis sie endlich etwas aussagen. Sie untersuchen zusätzliche Endpunkte, manipulieren Gruppenvergleiche und führen zahlreiche Subgruppenanalysen sowie wiederholte Zwischenanalysen durch. Die Schwierigkeiten zeigen sich gewöhnlich in der Auswertungsphase, wenn zusätzliche, nicht geplante Analysen durchgeführt wurden. Aus einer Studie können sich potenziell Tausende von Vergleiche ergeben, womit man allein zufallsbedingt viele signifikante Ergebnisse zu erwarten hat. Um dem zu begegnen, empfehlen manche Statistiker Adjustierungen, durch die häufig aber leider mehr Probleme entstehen als gelöst werden [1].

Multiplizitätsprobleme haben mehrere Ursachen. Hier werden wir uns mit mehrfachen Endpunkten sowie mehrfachen Behandlungen befassen. In einem zweiten Artikel [2] wird es um Subgruppenanalysen und Zwischenauswertungen gehen. Die Meinungen über Multiplizität gehen stark auseinander und sind komplexer Natur [3–6]. Welche Position zum Umgang mit Multiplizität man auch beziehen mag, jeder Vorschlag wird seine Kritiker finden (Kasten 1). Multiplizitätsfragen lösen hitzige Debatten aus [10].

Kasten 1: Divergierende Ansichten über statistische Adjustierungen für Multiplizität.

Manche Statistiker bevorzugen Adjustierungen für multiple Vergleiche, während andere sie ablehnen. „Verschiedene neuere Publikationen zeigen, dass die Debatte über multiple Vergleiche nach wie vor rege geführt wird. Meiner Beobachtung nach ist es sehr schwierig, Ansichten wie die folgenden miteinander in Einklang zu bringen …" [7]:

- „Bei multiplen Vergleichen ist die Durchführung von Adjustierungen nicht erforderlich." [4]
- „Bonferroni-Adjustierungen sind bestenfalls unnötig; schlimmstenfalls wirken sie sich auf eine solide statistische Schlussfolgerung nachteilig aus." [1]
- „… Der Fehler 1. Art wächst mit jedem Hypothesentest weiter an und muss kontrolliert werden." [8]
- „In wissenschaftlichen Arbeiten zu epidemiologischen und Public-Health-Themen, die multiple statistische Tests beinhalten, sollten die Methoden zur Identifizierung und Korrektur von Fehlern 1. Art dargelegt werden." [9]

Das Problem

Im Wesentlichen gibt es zwei Gründe dafür, warum Multiplizität sowohl beim Wissenschaftler als auch beim Leser gleichermaßen für Ärger sorgt. Zum einen sollten Studienautoren alle analytischen Vergleiche berichten, die sie durchgeführt haben. Leider wird die vollständige Auswertung der Daten zuweilen aber verschwiegen. Das erschwert es dem Leser, die betreffenden Studienergebnisse zu verstehen. Zum anderen kommen aber, wenn alle vorgenommenen Vergleiche korrekt berichtet werden, Statistiker mit ihren statistischen Adjustierungen an, die den multiplen Vergleichen Rechnung tragen sollen. Studienautoren müssen also wissen, ob sie derlei Adjustierungen vornehmen sollten, und die Leser, ob sie sie erwarten dürfen.

Multiplizität kann den Gesamtfehler bei der Signifikanzprüfung erhöhen. Unter der Hypothese, dass zwischen zwei Faktoren kein Zusammenhang besteht, gibt der Fehler 1. Art (oder α-Fehler) die Wahrscheinlichkeit an, mit der der aus den vorliegenden Daten beobachtete Zusammenhang zufällig ist. Er gibt dem Leser die Wahrscheinlichkeit einer falsch-positiven Schlussfolgerung an [11]. Ein Problem tritt dann auf, wenn mehrere voneinander unabhängige Assoziationen auf Signifikanz getestet werden. Wenn k die Anzahl der Vergleiche ist, dann ist die Wahrscheinlichkeit, dass mindestens eine Assoziation signifikant ist, $1 - (1 - \alpha)^k$. Häufig wird in der medizinischen Forschung für α ein Wert von 0,05 gewählt. Werden nun zehn voneinander unabhängige Assoziationen getestet, ausgehend von der globalen Nullhypothese, dass bei allen zehn kein Zusammenhang besteht, dann liegt die Wahrscheinlichkeit für mindestens ein signifikantes Ergebnis bei 0,40, nämlich $[1 - (1 - 0{,}05)^{10}]$. Anders ausgedrückt: Die kumulative Wahrscheinlichkeit für mindestens ein falsch-positives Ergebnis beträgt bei zehn Vergleichen 40%. Aber: Die Wahrscheinlichkeit eines falsch-positiven Ergebnisses bleibt für jeden einzelnen Vergleich trotzdem bei 0,05 (5%), ob nun eine oder eine Million Hypothesen getestet werden [4].

Vorschlag für eine statistische Lösung des Problems

Die meisten Statistiker würden dazu raten, die Anzahl der Vergleiche zu reduzieren, um das Problem des multiplen Testens zu lösen. Bei sehr vielen Tests schlagen manche Statistiker aber auch Adjustierungen vor, sodass die Gesamtwahrscheinlichkeit eines falsch-positiven Ergebnisses nach der Durchführung von *k* Vergleichen in der Studie dem Wert α entspricht. Gewöhnlich wird ein solches Vorgehen Bonferroni zugeschrieben und einfach festgestellt, dass man, um Vergleiche in einer Studie zum Signifikanzniveau α zu testen, alle Vergleiche auf dem Niveau α/k und nicht auf dem Niveau α durchführen sollte [5, 12]. Für $\alpha = 0{,}05$ müsste bei zehn Vergleichen jeder Test auf dem Niveau 0,005 signifikant sein. Demzufolge behalten manche Studienautoren jeweils denselben Schwellenwert α bei, multiplizieren aber jeden beobachteten *p*-Wert mit *k* [10, 13]. Bei zehn Vergleichen würde demnach ein in einer Studie beobachteter *p*-Wert von 0,02 adjustiert $p = 0{,}20$ ergeben. Zu beachten ist dabei jedoch, dass die Bonferroni-Korrektur den Fehler 2. Art (oder β-Fehler) vergrößert und dadurch die statistische Power verringert [1].

Allerdings bezieht sich die Bonferroni-Adjustierung in der Regel auf die falsche Hypothese [1, 6]. Sie geht von der globalen Nullhypothese aus, mit der (einfach ausgedrückt) geprüft wird, ob zwei Gruppen hinsichtlich aller untersuchten Primärendpunkte identisch sind, im Gegensatz zur Alternativhypothese, nach der ein Effekt für mindestens einen dieser Endpunkte vorliegt. In der medizinischen Forschung ist diese Frage aber gewöhnlich irrelevant. Klinisch gesehen entspräche dies etwa „… dem Fall eines Arztes, der für einen Patienten 20 verschiedene Laboruntersuchungen anordnet, nur um zu erfahren, dass einige dieser Tests Auffälligkeiten zeigen, ohne weitere Einzelheiten." [1]

Rothman schreibt dazu: „Die globale Nullhypothese zu vertreten heißt im Grunde, den Glauben an die reale Welt aufzugeben und damit die Grundfesten des Empirismus in Frage zu stellen." [4].

Die Arzneimittelzulassung, die nach klaren, dichotomen Antworten verlangt, ist ein starker Motor für die Beschäftigung mit der Adjustierung für Multiplizität. Adjustierungen passen zum Paradigma der für die Zulassung von Medikamenten erforderlichen Hypothesentestung – nämlich Zulassung oder Nichtzulassung. In der Mehrzahl der medizinwissenschaftlichen Publikationen hingegen befürworten wir eher die Angabe von Intervallschätzungen zur Bestimmung von Effekten (etwa relative Risiken mit Konfidenzintervallen) als das reine Hypothesentesten (nur *p*-Werte) [14]. Wir meinen sogar, dass der Wunsch nach einer Entscheidungsfindung in einem Großteil der medizinischen Forschung der Adjustierung für Multiplizität entgegensteht.

Mehrfache Endpunkte

Auch wenn das ideale Vorgehen bei der Planung und Auswertung von randomisierten, kontrollierten Studien auf der Untersuchung eines Primärendpunktes beruht, wird üblicherweise mehr als nur ein Endpunkt untersucht. Ein besonders krasser Missbrauch im Zusammenhang mit Multiplizität ergibt sich aus dem Ausheben oder Ausbaggern von Daten (Data Dredging), das sich hinter den Kulissen abspielt und nicht berichtet wird. Studienautoren analysieren viele Endpunkte, geben aber nur die günstigen signifikanten Vergleiche an. Wenn aber nur die signifikanten Vergleiche berichtet werden, dann gilt ein solches Vorgehen, falls es unabsichtlich geschieht, als unwissenschaftlich. Erfolgt es jedoch vorsätzlich, handelt es sich schlicht um Betrug. „Die *Post-hoc-Auswahl* des Endpunktes mit den signifikantesten Behandlungsunterschieden ist ein betrügerischer Trick, der stets dazu führt, dass ein Behandlungsunterschied überbetont wird." [13] Solchen Täuschungsmanövern muss Einhalt geboten werden.

Die Anzahl der primär zu untersuchenden Endpunkte sollte begrenzt und der oder die Primärendpunkte *a priori* im Studienprotokoll festgelegt werden. Ei-

Z. ärztl. Fortbild. Qual. Gesundh.wes. (2006) 100; 617–623
http://www.elsevier.de/zaefq

ne Fokussierung der jeweiligen Studie erleichtert ihre Implementierung und erhöht die Glaubwürdigkeit der Ergebnisse. Darüber hinaus sollte bei der Analyse das Studienprotokoll befolgt werden. Abweichungen vom Studienprotokoll zum Zwecke eines *Data Dredging* sind verzeihlich, sollten aber deutlich als explorativ gekennzeichnet und vollständig berichtet werden. Leider enthalten Studienberichte häufig die Untersuchung von Endpunkten, die im Studienprotokoll nicht erwähnt sind, während im Protokoll geplante Primäranalysen einfach nicht mehr auftauchen [15]. Vorsichtsmaßnahmen, die sicherstellen, dass sich die Studienautoren an das Studienprotokoll gehalten haben (wie etwa die Rückverfolgung von akzeptierten Studienprotokollen der Zeitschrift *The Lancet* und die Forderung nach Protokollen für alle randomisierten, kontrollierten Studien), sind dabei zwar hilfreich, sinnvoller jedoch ist eine noch umfassendere Registrierung und Veröffentlichung von Studienprotokollen. Und schließlich müssen alle tatsächlich durchgeführten Vergleiche berichtet werden [16, 17].

Statistische Adjustierungen für mehrfache Endpunkte können die Auswertung der Untersuchungsdaten sabotieren. Nehmen wir beispielsweise einmal an, es wurde eine randomisierte, kontrollierte Studie durchgeführt, in der ein neues Antibiotikum zur Prävention von febriler Morbidität nach Hysterektomie im Vergleich zu einem Standardantibiotikum untersucht wurde. Als primäre Zielgröße wurde Fieber angegeben, und die Ergebnisse ließen eine Senkung der febrilen Morbidität um 50% erkennen [relatives Risiko 0,50 (95%-CI 0,25 bis 0,99); $p = 0{,}048$]. Beachten Sie das signifikante Ergebnis. Nun stellen Sie sich vor, dass zwei primäre Endpunkte festgelegt wurden, und zwar Wundinfektion und Fieber. Wie in klinischen Studien üblich, sind diese Endpunkte stark korreliert. Abgesehen von der Fieberreduktion um 50% wurde in der Studie auch eine Abnahme der Wundinfektionen um 52% beobachtet [0,48 (95%-CI 0,24 bis 0,97); $p = 0{,}041$]. Manche Statistiker vertreten nun die Meinung, für mehrfache Vergleiche sollte adjustiert werden, indem beispielsweise jeder *p*-Wert mit der Anzahl der durchgeführten Vergleiche multipliziert wird – in diesem Fall hieße das $0{,}048 \times 2 = 0{,}096$ und $0{,}041 \times 2 = 0{,}082$. Beide *p*-Werte werden auf $>0{,}05$ adjustiert; damit wäre das Ergebnis der Studie unbestimmt (negativ).

Erfahrene Kliniker sehen diese Ergebnisse allerdings ganz anders. Das erste Ergebnis für den Endpunkt Fieber wird durch das Ergebnis für den Endpunkt Wundinfektion nicht etwa abgeschwächt, sondern untermauert. Ärzte wissen, dass die beiden Endpunkte biologisch gesehen in einem engen Zusammenhang stehen. Der zusätzliche zweite Endpunkt (Wundinfektion) und die Beobachtung ähnlicher Ergebnisse verleiht der beobachteten Senkung der febrilen Morbidität Glaubwürdigkeit. Dass Adjustierungen das eigentliche Ergebnis konterkarieren, verstößt gegen die Gesetze der Logik [1]. Dies käme in etwa dem Beispiel eines Arztes gleich, der bei einem Patienten einen krankhaft erniedrigten Hämoglobinwert feststellt, diesen aber nicht für behandlungswürdig hält, weil gleichzeitig auch ein pathologischer Hämatokritwert gefunden wurde.

In der Tat gibt es Statistiker, die in dem oben erwähnten Beispiel keine formalen Adjustierungen für Multiplizität durchführen würden. Und selbst Statistiker, die solche Adjustierungen normalerweise befürworten, empfehlen, in bestimmten klinischen Entscheidungsfindungsszenarien davon abzusehen [3]. Gilt es als ein Behandlungseffekt, wenn alle oder die meisten (im Studienprotokoll festgelegten) Endpunkte signifikant sind, dann sei keine Adjustierung für multiple Endpunkte erforderlich [3].

Überdies handelt es sich bei der Bonferroni-Adjustierung, die bei multiplen Vergleichen am häufigsten empfohlen wird, bestenfalls um eine Überkorrektur. Wenn die Endpunkte miteinander in Beziehung stehen [3, 13], was üblicherweise der Fall ist, kann dies auf eine schwerwiegende Überkorrektur hinauslaufen. Eine Überkorrektur für *p*-Werte beeinträchtigt aber die Interpretation der Ergebnisse. Durch Adjustierung für multiple Vergleiche „wird das Interpretationsproblem mechanisiert und dadurch auch trivialisiert und der Wert eines großen Teils der in umfangreichen Datenmengen enthaltenen Information negiert." [4]. Klinische Erkenntnisse haben nach wie vor Relevanz. Deshalb sollte man sich auf die kleinstmögliche Anzahl klinisch sinnvoller Endpunkte konzentrieren, dann aber die Ergebnisse aller untersuchten Endpunkte darlegen. Bei mehr als nur einem primären Endpunkt sollte diskutiert werden, ob zusätzliche Endpunkte die Kernresultate untermauern oder aber beeinträchtigen. Die formale Adjustierung für Multiplizität ist der Interpretation von Studienergebnissen eher hinderlich als förderlich.

Kombinierte Endpunkte

Kombinierte Endpunkte (Zielgrößen) können die mit Multiplizität verbundenen Bedenken abschwächen [18]. Ein kombiniertes Zielereignis tritt dann ein, wenn mindestens eines der prospektiv definierten Zielereignisse eintritt, aus denen sich der kombinierte Endpunkt zusammensetzt. So ergäbe sich etwa ein kombinierter kardiovaskulärer Endpunkt, wenn ein Myokardinfarkt, ein Schlaganfall oder ein kardiovaskulär bedingter Tod eintritt. Legt man den kombinierten Endpunkt *a priori* als den primären Endpunkt fest, dann erübrigen sich die mit dem Testen der Einzelkomponenten verbundenen multiplen Vergleiche. Außerdem führen kombinierte Endpunkte im Allgemeinen zu hohen Ereignisraten und steigern dadurch die Power der Studie bzw. verringern den benötigten Stichprobenumfang. Die häufige Anwendung kombinierter Endpunkte kann daher nicht überraschen [18].

Manchmal kommt es jedoch zu Interpretationsproblemen. So führte beispielsweise die Gabe von Aspirin in einer Studie zur Reduktion des oben erwähnten kombinierten Endpunktes kardiovaskulärer Ereignisse (Myokardinfarkt, Schlaganfall oder kardiovaskulär bedingter Tod) um 18% (relatives Risiko 0,82; 95%-CI 0,70 bis 0,96) – also ein scheinbar sehr vorteilhaftes Ergebnis [19]. Bei näherer Betrachtung

der einzelnen Komponenten dieses Endpunktes ergab sich jedoch eine Abnahme der Myokardinfarkte um 44%, eine Erhöhung der Schlaganfälle um 22% und hinsichtlich der kardiovaskulär bedingten Todesfälle nahezu keine Wirkung. Angesichts des fehlenden Benefits im Hinblick auf die vergleichsweise wichtigeren Zielgrößen Tod und Schlaganfall erscheint die Reduktion von 18% also eher unbedeutend zu sein [19]. Oftmals mangelt es kombinierten Endpunkten an klinischer Relevanz [20]. Sie adressieren das Problem der Multiplizität und sorgen gewöhnlich für statistische Effizienz um den Preis von Interpretationsproblemen.

Mehrfache Behandlungen (mehrarmige Studien)

Multiplizität, die sich aus mehrfachen Behandlungen ergibt, ist einfacher zu handhaben als Multiplizität aufgrund von multiplen Endpunkten. Zum einen können multiple Tests durch einen globalen, die Vergleichsgruppen übergreifenden Signifikanztest vermieden werden [13] – z. B. durch Vergleich von A *vs.* B *vs.* C in einer dreiarmigen Studie - oder durch Modellierung einer Dosis-Wirkungs-Beziehung [21]. Und zweitens, was vielleicht noch wichtiger ist, hat der Untersucher, wenn er viele Behandlungen vergleicht, seltener die Gelegenheit zum *Data Dredging*, ohne die Behandlungen auch alle anzugeben. Während es keine Probleme bereitet, bei der Datenanalyse mehr Endpunkte hinzuzufügen, ist es sehr viel schwieriger, zusätzliche Behandlungen in eine Studie aufzunehmen. Theoretisch könnte man eine Studie mit mehreren Behandlungsgruppen implementieren und dann aber nur die günstigen Gruppenvergleiche berichten; für ein solches Vorgehen gibt es allerdings in der Praxis nur wenige Belege. Wir gehen davon aus, dass der Leser eines Studienberichts normalerweise alle durchgeführten Behandlungen auch nachvollziehen kann. Und tatsächlich spielen mehrarmige Studien in der medizinischen Forschung eine wichtige Rolle (Kasten 2).

Kasten 2: Stellenwert mehrarmiger Studien in der medizinischen Forschung

Mehrarmige Studien sind in der medizinischen Literatur recht häufig anzutreffen. Die Suche nach parallel angelegten randomisierten, kontrollierten Studien, die im Jahre 2000 in *PubMed* indexiert waren, ergab, dass 25% der Studien mehr als zwei Behandlungsarme aufwiesen. In 62% der Fälle handelte es sich um dreiarmige Studien, 26% hatten vier und 12% mehr als vier Studienarme (D.G. Altman, persönliche Mitteilung).

In Lehrbüchern über klinische Studien werden überwiegend zweiarmige Studien behandelt. Überdies haben sich anerkannte Wissenschaftler vehement gegen Studien mit mehr als zwei Armen ausgesprochen: „Ein positives Ergebnis ist wahrscheinlicher und ein Null-Ergebnis informativer, wenn der Hauptvergleich aus nur zwei Therapien besteht, die sich möglichst stark voneinander unterscheiden." [22] Das Argument gegen mehrarmige Studien gründet sich hauptsächlich auf die statistische Trennschärfe (Power) einer Studie. Veröffentlichte Studien weisen üblicherweise eine zu geringe statistische Power auf [23]. Bei einer endlichen Anzahl potenzieller Studienteilnehmer muss man davon ausgehen, dass die Hinzufügung weiterer Studienarme lediglich zur Verwässerung der statistischen Power führt. Auch wenn wir diesem Argument prinzipiell zustimmen, gibt es jedoch bestimmte Umstände, unter denen mehrarmige Studien möglicherweise aber nicht nur attraktiv, sondern auch effizienter sind.

Stellen Sie sich dazu beispielsweise einen Fall vor, in dem eine Standardbehandlung existiert und zwei neue, potenziell wirksame Therapien entwickelt wurden. Ein zweiarmiger Ansatz verlangt den Vergleich einer der beiden neuen Therapien mit der Standardbehandlung und dann wahrscheinlich eine zweite Studie, in der die andere neue Therapie mit einer Gruppe aus der ersten Studie verglichen wird. Alles in allem wären Gesamtstudiengröße und -kosten bei diesem sequenziellen zweiarmigen Vorgehen höher als bei einer mehrarmigen Studie. Manchmal sind mehrarmige Studien also durchaus sinnvoll. Außerdem rufen mehrarmige Studien nicht zwangsläufig methodische Bedenken hervor. Ebenso wie in zweiarmigen kann Selektionsbias auch in mehrarmigen Studien ausgeschaltet werden. Und auch wenn die Durchführung und Auswertung von mehrarmigen Studien sich komplexer gestaltet, führt diese Komplexität häufig doch auch zu einem entsprechend höheren Informationsgewinn.

Ganz so vorteilhaft stellt sich die Lage in Wirklichkeit aber nicht dar. Was der Leser eines Zeitschriftenartikels möglicherweise nämlich nicht zu sehen bekommt, sind all die verschiedenen Vergleiche zwischen den Therapiegruppen. Bei einer dreiarmigen Studie beispielsweise ergeben sich mindestens sieben mögliche Analysen (Abb.). Bei mehr als drei Armen steigt die Zahl der potenziellen Vergleiche sogar explosionsartig an. Die Zahl der beabsichtigten Vergleiche sollte also a priori spezifiziert werden.

Bei mehrarmigen Studien wird, wie bereits erwähnt, häufig angeraten, einen globalen, alle Behandlungen übergreifenden Test durchzuführen. Manche Methodiker sind jedoch der Meinung, dass solche Tests nur von begrenztem Nutzen sind, weil man damit nicht feststellen kann, welche Therapien sich unterscheiden, und weil ihre statistische Power zum Nachweis echter Unterschiede begrenzt ist [13]. Viele mehrarmige Studien sind für den direkten Vergleich mit Kontrollen ausgelegt [13]. Deshalb sollten Untersucher die beabsichtigten Vergleiche planen, ihre Anzahl begrenzen und sie im Protokoll dokumentieren.

Statistische Korrekturen für multiple Vergleiche sind in Studien mit mehreren Behandlungsgruppen nicht unbedingt erforderlich. Ähnlich wie bei dem oben angeführten Argument für multiple Endpunkte werden Ärzte im Allgemeinen feststellen, dass sich der Informationsgehalt einer Studie durch Hinzunahme einer weiteren Gruppe eher erhöht als verringert. Angenommen, in die bereits beschriebene randomisierte, kontrollierte Studie über den Vergleich eines neuen Antibiotikums mit der Standardtherapie zur Fieberprävention nach Hysterektomie wurde zusätzlich zur Gruppe mit einer Dosis von 200 mg eine Behandlungsgruppe aufgenommen, die 300 mg desselben Antibiotikums erhält. Die Ergebnisse zeigen eine Reduktion von 40% für die 200-mg-Dosis (relatives Risiko 0,60; 95%-CI 0,37 bis 0,98; $p = 0{,}044$). Man beachte, dass es sich hierbei um ein signifikantes Ergebnis handelt. Wie zu erwarten war, erhielt man für die 300-mg-Dosis mit einer Abnahme der Fieberfäl-

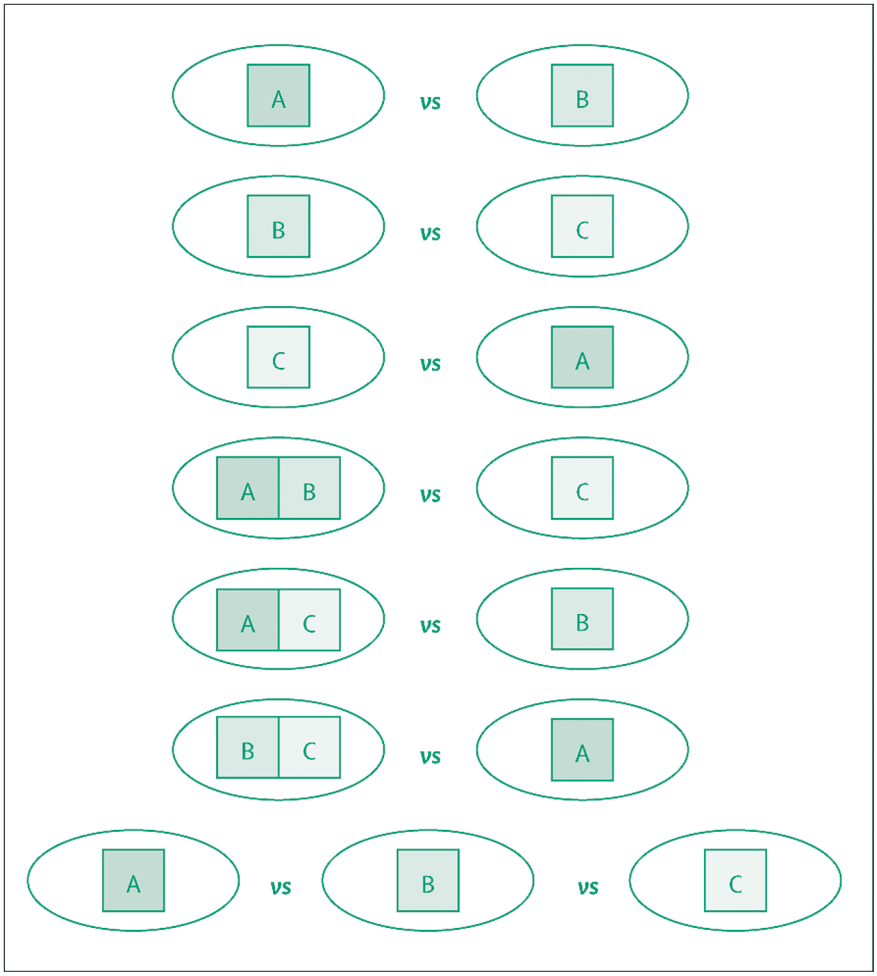

Abb. 1. Dreiarmige Studie mit mindestens sieben Vergleichsmöglichkeiten.

le um 45% ein ähnliches Ergebnis (0,55; 95%-CI 0,31 bis 0,98; p = 0,041). Bei einer einfachen Adjustierung für multiple Vergleiche wird jeder einzelne p-Wert mit der Anzahl der durchgeführten Vergleiche multipliziert – d. h. 0,044 × 2 = 0,088 und 0,041 × 2 = 0,082. Durch die Adjustierung werden die Effekte auf dem 0,05-Niveau also nichtsignifikant und damit unbestimmt (negativ).

Auch in diesem Fall werden die Ergebnisse von den Untersuchern ganz anders interpretiert. Im Hinblick auf den Endpunkt Fieber wird das Ergebnis der 200-mg-Gruppe durch das Resultat der 300-mg-Gruppe nicht etwa abgeschwächt, sondern untermauert. Auch aus biologischer Sicht würden Ärzte mit einem solchen Ergebnis rechnen. Einer Adjustierung, die diese signifikanten Ergebnisse zunichte macht, würden sie daher ernsthaft misstrauen.

Die Auswertung der Studienergebnisse wird durch die Adjustierung von p-Werten, vor allem bei vergleichbaren Behandlungen, also nicht gerade unterstützt.

Gelegentlich werden bei mehreren Behandlungen Tests in einer nach Priorität geordneten Reihenfolge durchgeführt [24]. So könnte man sich etwa entschließen, den Vergleich zwischen der Standardtherapie und der 300-mg-Dosis des neuen Antibiotikums als Test mit erster Priorität durchzuführen und mit dem 200-mg-Vergleich nur fortzufahren, wenn der erste Vergleich signifikant ausfällt. Solche Testprozeduren tragen dem Problem des multiplen Testens Rechnung, ohne dass man eine Adjustierung durchführen muss [24]. Auch hier gilt wieder, dass formale Adjustierungen für Multiplizität eher Komplikationen verursachen als zur Klärung beitragen.

Stellenwert von Adjustierungen für Multiplizität

Manchmal kommt man allerdings nicht ohne eine formale Adjustierung für multiple Vergleiche aus. Ein einleuchtendes Beispiel dafür wären etwa bestimmte Entscheidungskriterien im Zusammenhang mit der Einreichung eines Arzneimittelzulassungsantrages bei einer Zulassungsbehörde: So sollte man etwa dann für Multiplizität korrigieren, wenn der Sponsor mehr als einen Primärendpunkt spezifiziert und vorschlägt, von einem Behandlungserfolg auszugehen, wenn einer oder mehrere dieser Endpunkte sich als signifikant erweisen [3]. Dies lässt sich auf alle Situationen übertragen, in denen man dann einen Behandlungseffekt annimmt, wenn mindestens einer von mehreren Endpunkten ein positives Ergebnis zeigt.

Adjustierungen können auch in einer mehrarmigen Studie angezeigt sein, in der eine Auswertung „jeder gegen jeden" geplant ist, etwa in einer vierarmigen Studie mit den Behandlungen A, B, C und D, in der man von einem Behandlungseffekt für A spricht, wenn einer der folgenden Vergleiche signifikante Ergebnisse liefert: A versus B, A versus C, A versus D, A versus B + C, A versus B + D, A versus C + D oder A versus B + C + D. Hier wäre es wohl das Beste, für multiple Vergleiche zu adjustieren.

Erscheint eine Adjustierung für Multiplizität angeraten, kann man im Allgemeinen davon ausgehen, dass es sich um schlecht und diffus angelegte Studien handelt. Aber auch durch eine Adjustierung lässt sich die Glaubwürdigkeit einer solchen Studie nur zum Teil retten. Und selbst wenn diese sich als angemessen erweist, wird die Implementierung zum Problem.

Im Allgemeinen wird – meist wegen ihrer Einfachheit – die Bonferroni-Korrektur empfohlen. Manchmal lassen sich mit anderen Adjustierungsstrategien jedoch bessere Ergebnisse erzielen [3, 25]. Simulationsexperimente zeigen in Abhängigkeit von der Korrelation zwischen den Endpunkten im Hinblick auf Fehler 1. Art und die Power der verschiedenen Strategien zur Adjustierung

für Multiplizität ein breites Schwankungsspektrum [3]. Zwar sind diese vergleichenden Wertungen hilfreich, doch bleibt es schwierig, zu eindeutigen Aussagen zu gelangen. Meist liefern Adjustierungen nur grobe Antworten.

Worauf der Leser achten sollte

Der Leser sollte davon ausgehen können, dass alle in einer Studie untersuchten Endpunkte und Therapievergleiche im Studienbericht Erwähnung finden. Ob auch wirklich alle Endpunkte bzw. alle Vergleiche angegeben wurden, lässt sich aber normalerweise nur schwer nachprüfen. Von Vorteil, meist jedoch nicht möglich ist der Zugang zum jeweiligen Studienprotokoll. Aus diesem Grund drängen wir darauf, den Zugriff auf Studienprotokolle zu verbessern. Durch eine mangelhafte, unvollständige Berichterstattung wird dem Leser die von den Studienautoren durchgeführte komplette Datenanalyse häufig jedoch vorenthalten. Die Niederlegung von Studienergebnissen gemäß dem CONSORT-Statement beugt solchen Problemen vor [16, 17].

Der Leser sollte erwarten dürfen, dass der oder die Primärendpunkte spezifiziert und weitere Analysen als explorativ kenntlich gemacht werden. Er sollte anstelle von direkten Erklärungen auch nach indirekten Hinweisen suchen. Wenn der primäre Endpunkt unklar bleibt, dann haben die Autoren hoffentlich eine Analyse zur statistischen Power durchgeführt, aus der sich der primäre Endpunkt ableiten lässt.

Ferner darf der Leser auch eine Erklärung erwarten, wenn in einer Studie multiple Vergleiche angestellt werden. Wenn es die Autoren übertreiben und Ergebnisse für 15 Endpunkte berichten, von denen nur einer signifikant ist, ist entsprechende Vorsicht geboten. Und wenn multiple Vergleiche multiple Effekte erkennen lassen, sollten sich die Autoren mit der internen Konsistenz der Ergebnisse auseinandergesetzt haben. Am wichtigsten ist jedoch, dass eine transparente Berichterstattung aller Vergleiche den Leser in die Lage versetzt, seine eigenen Interpretationen anzustellen.

Wird in einem Studienbericht ein kombinierter Endpunkt spezifiziert, sollten dessen einzelne Komponenten den bekannten pathophysiologischen Merkmalen der betreffenden Erkrankung entsprechen. Ein kombinierter Endpunkt sollte allerdings als Ganzes interpretiert und nicht zum Nachweis der Wirksamkeit seiner einzelnen Komponenten benutzt werden. Die Komponenten sollten als sekundäre Zielgrößen spezifiziert und zusätzlich zu den Ergebnissen der Primäranalyse berichtet werden [18].

Generell braucht der Leser nicht mit Korrekturen für multiples Testen zu rechnen. In den meisten Studien mangelt es solchen Adjustierungen an Substanz, und sie sind wenig hilfreich. Eine Ausnahme bilden vielleicht Arbeiten, in denen sich die Beweisführung ausschließlich darauf stützt, dass wenigstens einer der Primärendpunkte signifikant ist, also tatsächlich auf die Überprüfung der globalen Nullhypothese. Durch Adjustierung für Multiplizität lassen sich solche Analysen bis zu einem gewissen Grade retten.

Erklärung zum Ausschluss von Interessenkonflikten

Hiermit erklären wir, dass keinerlei Interessenkonflikte vorliegen.

Danksagungen

Wir danken David L. Sackett, Douglas G. Altman und Willard Cates für ihre hilfreichen Kommentare zu einer früheren Version dieses Artikels.

Literatur

[1] Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998;316: 1236–8.
[2] Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. Lancet 2005;365: 1657–61.
[3] Sankoh AJ, D'Agostino RB Sr, Huque MF. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple

endpoint issues. Stat Med 2003;22: 3133–50.
[4] Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1:43–6.
[5] Westfall P, Bretz F. Multiplicity in clinical trials: encyclopedia of biopharmaceutical statisitics, 2nd edn. New York: Marcel Dekker, 666–673 (2003).
[6] Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. Am J Epidemiol 1995;142:904–8.
[7] Altman DG. Statistics in medical journals: some recent trends. Stat Med 2000;19:3275–89.
[8] Moye LA. P-value interpretation and alpha allocation in clinical trials. Ann Epidemiol 1998;8:351–7.
[9] Ottenbacher KJ. Quantitative evaluation of multiplicity in epidemiology and public health research. Am J Epidemiol 1998;147:615–9.
[10] Altman DG. Practical statistics for medical research. London: Chapman and Hall (1991).
[11] Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005;365: 1348–53.
[12] Friedman L, Furberg C, DeMets D. Fundamentals of clinical trials. St Louis: Mosby (1996).
[13] Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley (1983).
[14] Sterne JA, Davey Smith G. Sifting the evidence: what's wrong with significance tests? BMJ 2001;322:226–31.
[15] Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291:2457–65.
[16] Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports or parallel group trials. Lancet 2001;357:1191–4.
[17] Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663–94.
[18] Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA 2003;289:2554–9.
[19] Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. Control Clin Trials 1997;18: 530–45.
[20] Meinert CL. Clinical trials: design, conduct, and analysis. New York: Oxford University Press (1986).
[21] Senn S. Statistical issues in drug development. Chichester: John Wiley and Sons (1997).
[22] Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient I: introduction and design. Br J Cancer 1976;34:585–612.
[23] Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122–4.
[24] Bauer P, Chi G, Geller N, et al. Industry, government, and academic panel discussion on multiple comparisons in a "real" phase three clinical trial. J Biopharm Stat 2003;13:691–701.
[25] Hsu JC. Multiple comparisons: theory and methods. New York: Chapman and Hall (1996).

Korrespondenzadresse:
Dr. Kenneth F. Schulz
E-Mail: KSchulz@fhi.org

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.