Available online at www.sciencedirect.com
ScienceDirect
ZaeFQ
www.elsevier.de/zaefq
Z.ärztl. Fortbild. Qual.Gesundh.wes. (ZaeFQ) 101 (2007) 51–58

Methodik

Reihe Epidemiologie 5
Multiplizität in randomisierten Studien II: Subgruppenanalysen und Zwischenauswertungen

Kenneth F. Schulz, David A. Grimes

Family Health International, PO Box 13950, Research Triangle Park, NC 27709 USA

Aus: Lancet (2005) ***365****: 1657–1661*
(Originaltitel: „Multiplicity in randomised trials II: subgroup and interim analyses")

Subgruppenanalysen können schwerwiegende, durch Multiplizität bedingte Probleme verursachen. Wenn genügend viele Subgruppen getestet werden, erhält man wahrscheinlich schon rein zufallsbedingt mindestens ein falsch-positives Ergebnis. Mitunter führen Untersucher zahlreiche Analysen durch, berichten aber nur die signifikanten Effekte. Das führt zu Verzerrungen in der medizinischen Literatur. Im Allgemeinen raten wir von Subgruppenanalysen ab. Wenn sie jedoch nötig sind, dann sollte man statistische Interaktionstests durchführen, anstatt die einzelnen Subgruppen getrennt voneinander auszuwerten. Zwischenauswertungen sind nicht zu vermeiden, wenn die Überwachung der Daten (Datenmonitoring) angebracht ist. Allerdings führt wiederholtes Testen bei jeder dieser Zwischenauswertungen zu Bedenken wegen der damit verbundenen Multiplizitätsprobleme, und wenn dieser Multiplizität nicht Rechnung getragen wird, schnellt die Irrtumswahrscheinlichkeit 1. Art (Rate falsch-positiver Ergebnisse) in die Höhe. Hier kommen statistische Abbruch- oder Stoppregeln ins Spiel. Durch die einfach anzuwendenden gruppensequenziellen Stoppregeln von O'Brien & Fleming und von Peto bleiben das angestrebte Signifikanzniveau und die statistische Trennschärfe (Power) erhalten. Beide Verfahren setzen für die Zwischenauswertungen strenge Kriterien (niedrige nominelle *p*-Werte) an. Die Implementierung einer Studie unter solchen Abbruchregeln ähnelt der einer konventionellen Studie, mit der einzigen Ausnahme, dass die Studie frühzeitig beendet werden kann, wenn sich eine der untersuchten Therapien als deutlich überlegen erweisen sollte. Untersucher und Leser müssen sich jedoch darüber im Klaren sein, dass die geschätzten Therapieeffekte bei einem frühzeitigen Abbruch der Studie zufallsbedingt eine Tendenz zu übertrieben hohen Werten zeigen.

Subgruppenanalysen üben einen besonderen Reiz aus. Sie scheinen logisch und intuitiv zu sein – und machen Untersuchern wie auch Lesern sogar Spaß. Diese tückische Attraktivität verursacht jedoch relevante Probleme. Multiplizität gepaart mit Naivität verleitet bei der Durchführung von Studien und der Präsentation ihrer Ergebnisse zu Interpretationsfehlern. Die aus Subgruppen abgeleiteten Behandlungseffekte, die in vielen Studienberichten auftauchen, können sich aber als illusorisch erweisen.

E-Mail: k.schulz@fhi.org (K.F. Schulz)

Im Gegensatz dazu sind Zwischenauswertungen nicht vermeidbar, wenn die Überwachung der Daten erforderlich ist. Auch kann man für diese Zwischenauswertungen nicht die normalen statistischen Vorgehensweisen wählen. Um das Monitoring von Studienergebnissen zu unterstützen, müssen statistische Stoppregeln angewendet werden, im Wesentlichen statistische Adjustierungen, die eher Warnungen als Abbruchregeln darstellen. Leider verwirren solche Verfahren sowohl Untersucher als auch Leser gleichermaßen. Die Statistik erweist sich schon oft genug als verwirrend, ohne dass Abbruchregeln auch noch Komplikationen zweiter Ordnung ins Spiel bringen.

Durch Subgruppen- und Zwischenanalysen verursachte Probleme der Multiplizität stellen uns vor ähnliche Schwierigkeiten, wie sie im Zusammenhang mit multiplen Endpunkten und dem Vergleich mehrerer Therapiegruppen auftreten [1]. Häufig baggern Untersucher ihre Daten richtiggehend aus, indem sie zahlreiche Subgruppenanalysen durchführen und wiederholt Zwischenauswertungen vornehmen. Außerdem führen sie zuweilen auch ungeplante Subgruppen- und Zwischenanalysen durch. Und doch unterscheiden sich manche Lösungsansätze für Multiplizitätsprobleme, die durch Subgruppenanalysen und Zwischenauswertungen entstehen, von den Lösungsvorschlägen für Probleme, die durch multiple Endpunkte und multiple Therapievergleiche verursacht werden.

Subgruppenanalysen

Willkürlich durchgeführte Subgruppenanalysen rufen wegen der damit verbundenen Multiplizitätsprobleme schwerwiegende Bedenken hervor. Diese Probleme finden ihren Widerhall in der gesamten medizinwissenschaftlichen Literatur. Denn trotz zahlreicher Warnungen [2] gibt es noch immer Untersucher, die stur auf der Durchführung allzu vieler Subgruppenanalysen beharren.

Zu Beginn einer Studie werden auf der Grundlage der Patientencharakteristika bestimmte Subgruppen von Teilnehmern definiert. Dann führt man Analysen durch, um festzustellen, ob sich in diesen Subgruppen unterschiedliche Therapieeffekte beobachten lassen. Die Hauptschwierigkeiten rühren daher, dass die Untersucher auf jede der untersuchten Subgruppen statistische Tests anwenden. Die Auswertung mehrerer Subgruppen in Kombination mit mehreren Zielgrößen (Outcomes) lässt die Anzahl der statistischen Tests übermäßig anwachsen.

Die Suche nach positiven Subgruppeneffekten in Studien ohne nachweisbaren Gesamteffekt (Data Dredging) könnte der Antrieb für ein solches Vorgehen sein. Denn wenn man genügend viele Subgruppen testet, dann erhält man allein zufallsbedingt falsch-positive Ergebnisse.

„Die Antwort auf eine randomisierte, kontrollierte Studie, in der sich die eigenen Annahmen nicht bestätigen, ist nicht etwa die Durchführung mehrerer Subgruppenanalysen, bis man endlich seine Annahmen bestätigt findet. Die Antwort darauf lautet vielmehr, dass man seine Annahmen gründlich überdenken sollte." [3]

Ähnlich kann auch das Testen von Subgruppen in Studien *mit* einem eindeutigen Gesamteffekt aufgrund von Zufall und mangelnder Studienpower zu falsch-negativen Ergebnissen führen.

In der Zeitschrift *The Lancet* wurde dazu ein anschauliches Beispiel veröffentlicht [4]. Aspirin hatte hinsichtlich der Prävention des plötzlichen Herztodes nach Myokardinfarkt eine starke positive Wirkung gezeigt ($p<0{,}00001$, mit schmalem Konfidenzintervall). Die Herausgeber drängten die Studienautoren dazu, nahezu 40 Subgruppenanalysen zu unternehmen [2]. Die Autoren stimmten diesem Ansinnen widerstrebend zu, stellten aber die Bedingung, dass sie eine eigene Subgruppenanalyse vorlegen durften, mit der sie die Unzuverlässigkeit solcher Subgruppenauswertungen unter Beweis stellen wollten. Sie wiesen nach, dass sich bei den Studienteilnehmern mit dem astrologischen Sternzeichen Zwilling oder Waage ein geringfügig unerwünschter Effekt auf die durch Aspirin bedingte Mortalität beobachten ließ (Anstieg um 9%, Standardabweichung (SD) 13; nichtsignifikant), während den unter den übrigen Sternzeichen geborenen Teilnehmern eine auffallend vorteilhafte Wirkung zuteil wurde (Reduktion um 28%, SD 5; $p<0{,}00001$) [4].

Entgegen anekdotischen Berichten einer Bestätigung seitens der Astrologen hat dieser zufällige „Sternzeichenbefund" bei den Medizinern nur geringes Interesse hervorgerufen. Aus ihren Subgruppenanalysen folgerten die Autoren: „All diese Subgruppenanalysen sollten vielleicht nicht als Evidenz dafür verstanden werden, wer von der Behandlung profitiert, sondern vielmehr als Beleg dafür, dass solche Analysen potenziell irreführend sind."

Wie diese und andere umsichtige Untersucher betonen, ist der zuverlässigste Effektschätzer für eine bestimmte Subgruppe der Gesamteffekt (im Wesentlichen alle Subgruppen zusammen genommen) und nicht der in der jeweiligen Subgruppe beobachtete Effekt [4,5]. Dieser Ansicht stimmen wir zu.

Eine korrekt durchgeführte Auswertung kann einen Großteil der durch Subgruppenanalysen entstehenden Multiplizitätsprobleme bereinigen. Unpassenderweise wenden Untersucher oftmals aber auf jede einzelne Subgruppe Tests an, und das öffnet Zufallsbefunden Tür und Tor. Wenn man beispielsweise die Teilnehmer nach ihrem Alter bei Studienbeginn in vier Kategorien aufteilt, ergeben sich allein für dieses Merkmal von vornherein schon vier Tests (Tabelle 1). In einer gut durchgeführten Analyse verwendet man einen statistischen Interaktionstest, mit dem sich feststellen lässt, ob der Behandlungseffekt für ein Outcome von der Subgruppe des jeweiligen Teilnehmers abhängt. Bei einem solchen Vorgehen wird nicht nur die richtige Frage getestet, vielmehr muss anstelle von vier auch nur ein Test durchgeführt werden, der im Wesentlichen das Multiplizitätsproblem adressiert. Untersucher stellen Interaktionstests immer wieder mit der Begründung in Frage, es mangele ihnen an statistischer Trennschärfe. Tatsächlich aber garantieren diese Interaktionstests die nötige Vorsicht. Sie erkennen, dass die von Subgruppen ausgehende Information

Tabelle 1. Wirkung eines neuen Antibiotikums im Vergleich zur Wirkung des Standardantibiotikums auf die febrile Morbidität in vier Altersgruppen und in der Gesamtsumme.

	Febrile Morbidität			Relatives Risiko (95%-CI)
	Ja	Nein	Gesamtsumme	
Alter 20–24 Jahre				
Neues Antibiotikum	11	84	95	1,4 (0,6 bis 3,2)
Standardantibiotikum	8	86	94	
Alter 25–29 Jahre				
Neues Antibiotikum	8	69	77	1,2 (0,4 bis 3,1)
Standardantibiotikum	7	72	79	
Alter 30–34 Jahre				
Neues Antibiotikum	3	48	51	0,3 (0,1 bis 0,9)
Standardantibiotikum	11	38	49	
Alter 35–39 Jahre				
Neues Antibiotikum	10	32	42	1,1 (0,5 bis 2,5)
Standardantibiotikum	9	33	42	
Gesamtsumme				
Neues Antibiotikum	32	233	265	0,9 (0,6 bis 1,4)
Standardantibiotikum	35	229	264	

Der Test auf statistische Interaktionen (Breslow-Day) ist nichtsignifikant ($p = 0,103$), was darauf schließen lässt, dass ein Subgruppenbefund in der Altersgruppe der 30–34-Jährigen zufallsbedingt ist. Ein solches Ergebnis wäre – wenn es in der Schlussfolgerung unangemessen hervorgehoben würde – ein Beispiel für einen überflüssigen Rettungsversuch per Subgruppenanalyse eines ansonsten unbestimmten (negativen) Studienergebnisses

begrenzt ist, und sind das wirksamste statistische Verfahren, um unangemessene Subgruppenergebnisse im Zaum zu halten, während sie Interaktionseffekte – soweit vorhanden – weiterhin nachweisen können [6,7].

Ein weiteres Problem im Zusammenhang mit Subgruppenanalysen ergibt sich daraus, dass Untersucher viele Analysen durchführen, aber nur solche mit signifikanten Ergebnissen berichten, was letzteren mehr Glaubwürdigkeit verleiht als ihnen zukommt. Ein solches Vorgehen ist irreführend und – falls es absichtlich geschieht – auch unethisch. Dies entspricht der Situation, die wir schon im Zusammenhang mit der Diskussion über multiple Endpunkte als Hauptproblem herausgearbeitet haben (siehe Z. ärztl. Fortbild. Qual. Gesundh.wes (2006), 100, 53–59).

Subgruppenanalysen stellen nach wie vor ein Problem in veröffentlichten Artikeln dar. In einem Review von 50 in allgemeinmedizinischen Fachzeitschriften (New England Journal of Medicine, The Lancet, JAMA und BMJ) publizierten Studienberichten wurde festgestellt, dass in 70% der Fälle Subgruppenanalysen präsentiert wurden [8]. In den Artikeln, in denen sich die Anzahl der Auswertungen ermitteln ließ, waren in nahezu 40% der Fälle mindestens sechs Subgruppenanalysen erfolgt – in einem Fall sogar 24. Statistische Interaktionstest waren in weniger als der Hälfte der Artikel durchgeführt worden. Überdies enthielten die Studienberichte keine Informationen darüber, ob die Subgruppenanalysen vorab geplant oder *post hoc* durchgeführt worden waren. Die Reviewautoren äußerten den Verdacht, dass „… einige Untersucher selektiv nur die interessanteren Subgruppenanalysen berichten und damit den Leser (wie auch uns) im Unklaren darüber lassen, wie viele weniger spannende Subgruppenanalysen sie sich angesehen und nicht erwähnt haben." [8] Leider wurde in der Mehrzahl der Studien, die über Subgruppenanalysen berichten, ein Subgruppenunterschied festgestellt, der in den Schlussfolgerungen besonders hervorgehoben wurde [8] – so viel zum Thema vorsichtige Interpretation!

Wir raten von Subgruppenanalysen ab. Bei korrekter Durchführung sind sie nicht unbedingt falsch. Manchmal sind sie biologisch betrachtet sinnvoll, oder sie werden von den (öffentlichen wie auch industriellen) Auftraggebern verlangt. Entscheidet man sich für die Durchführung solcher Analysen, dann sollte man sich auf die Primärzielgröße beschränken und die Anzahl der Subgruppen begrenzen. Geplante Subgruppenanalysen sollten vorab im Studienprotokoll festgelegt werden. Untersucher müssen alle Subgruppenanalysen, die sie vorgenommen haben, angeben, nicht bloß die mit signifikanten Ergebnissen. Wichtig ist, statistische Tests auf Interaktionen anzuwenden, um festzustellen, ob sich die Subgruppen hinsichtlich eines Behandlungseffekts unterscheiden, anstatt einzelne Tests in jeder Subgruppe durchzuführen. Durch ein solches Vorgehen lassen sich die Hauptbedenken gegen multiples Testen abschwächen. Die Schlussfolgerungen einer Studie sollten nur in seltenen Fällen von ihren Subgruppenanalysen beeinflusst sein.

„Subgruppenanalysen sind besonders anfällig für Überinterpretation, und man ist versucht, bei vielen Studien den Rat zu geben: „Lassen Sie die Finger davon!" (oder wenigstens zu sagen: „Glauben Sie nicht daran!"), doch dieser Rat läuft der menschlichen Natur wahrscheinlich zuwider." [8,9]

Methodiker halten sich mit ihrer Kritik an unzureichend durchgeführten Subgruppenanalysen zu sehr zurück. Solche methodischen Schwächen müssten viel stärker angeprangert werden.

Worauf Leser bei Subgruppenanalysen achten sollten

Der Leser sollte vor Studien auf der Hut sein, in denen Untersucher viele Subgruppenanalysen berichten, außer sie geben stichhaltige Gründe dafür an. Ferner sollte er sich vor Studien in Acht nehmen, in denen nur wenige Subgruppenanalysen angegeben sind. Es könnte nämlich sein, dass die Untersucher viele Subgruppenanalysen durchgeführt, sich aber nur die Rosinen, d.h. die interessanten und signifikanten Ergebnisse, herausgepickt haben. Die Unterdrückung von Analysen (Underreporting) gibt folglich Anlass zu der Behauptung, dass Studien mit wenigen Subgruppenanalysen sogar noch schlimmer sind als Studien mit vielen Subgruppenanalysen. Untersucher erscheinen glaubwürdiger, wenn sie versichern, dass sie alle durchgeführten Analysen auch angegeben haben. Ferner sollten Untersucher die nicht geplanten Subgruppenanalysen als Hypothesen generierend kennzeichnen, nicht als konfirmatorisch. Ergebnisse aus solchen Subgruppenanalysen sollten nicht in den Schlussfolgerungen auftauchen.

Bei Subgruppeneffekten sollten Leser Interaktionstests erwarten und Analysen, die auf innerhalb von Subgruppen durchgeführten Tests beruhen, nicht weiter beachten. Selbst bei einem signifikanten Interaktionstest sollte der Leser seine Interpretation der Ergebnisse auf die biologische Plausibilität, die vorherige Spezifikation der Analysen im Studienprotokoll und die statistische Aussagekraft der Informationen stützen. Adjustierungen für Multiplizität sind im Allgemeinen überflüssig, wenn die Untersucher Interaktionstests einsetzen. Aber angesichts des häufig zu beobachtenden ungehemmten Ausbaggerns von Studiendaten wiegt das Argument für statistische Adjustierungen im Fall von Subgruppenanalysen schwerer als bei multiplen Endpunkten. Überdies sind Adjustierungen für Multiplizität angemessen, wenn Untersucher keine Interaktionstests anwenden, für jede einzelne Subgruppe aber Tests angeben [10]. Die meisten Subgruppenergebnisse neigen dazu, die Wirklichkeit überspitzt darzustellen. Der Leser sollte sich besonders vor Untersuchern in Acht nehmen, die in Studien ohne einen Gesamtbehandlungseffekt einen Therapieeffekt in einer der Subgruppen besonders herausstellen [11]. Dabei handelt es sich gewöhnlich um überflüssige Rettungsversuche durch Subgruppen in Studien mit ansonsten unbestimmten (negativen) Ergebnissen (Tabelle 1) [8].

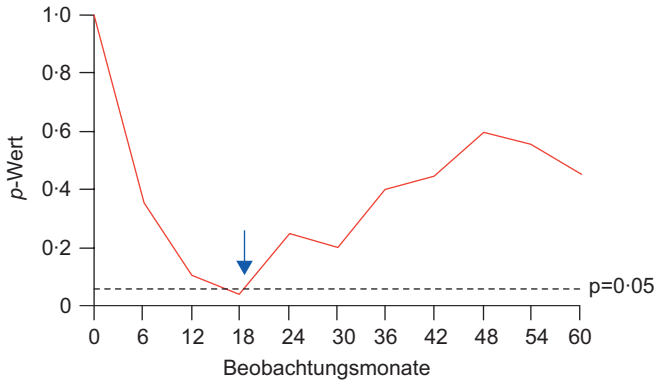
Zwischenauswertungen

Ein angemessenes Studienmonitoring umfasst mehr als nur statistische Warnregeln für den vorzeitigen Abbruch einer Studie. Tatsächlich kommt der Über- oder Unterlegenheit der untersuchten Behandlung eine wichtige Rolle zu. Allerdings könnten langsame Teilnehmerrekrutierung, mangelhafte Datenqualität, unzureichende Befolgung des Studienprotokolls, Ressourcenmängel, unakzeptable Nebenwirkungen, Betrug und das Bekanntwerden von Informationen, die die Studie irrelevant, unnötig oder unethisch erscheinen lassen, alle zum vorzeitigen Abbruch einer Studie führen. Der Entscheidungsprozess ist zweifellos komplex [12,13] und wird am besten in die Hände eines unabhängigen Datenüberwachungsgremiums (Data Monitoring Committee) gelegt. Die Aufgabe eines solchen Gremiums lässt sich mithilfe einer vorab festgelegten statistischen Abbruchregel bewältigen. Doch sind sich Untersucher und Leser dieser statistischen Probleme häufig gar nicht bewusst .

Die Akkumulation von Daten in Studien verleitet Untersucher dazu, die Daten für ihre Hauptzielgröße (schon vorab) auszuwerten. Wenn sie am Ende der Studie eine Irrtumswahrscheinlichkeit von $p<0{,}05$ anstreben, führen sie womöglich alle Zwischenauswertungen zum Signifikanzniveau von $\alpha = 0{,}05$ durch, und das ist falsch.

Zur Verdeutlichung dieses Problems mag die graphische Darstellung eines Beispiels beitragen. Ein Datenüberwachungsgremium führt 5 Jahre lang alle 6 Monate eine Zwischenauswertung der bis dahin angesammelten Studiendaten durch. Nach 18 Monaten rutscht die Analyse unter den p-Wert von 0,05 und erreicht dieses Signifikanzniveau danach nie wieder. Eine frühe Entscheidung des Gremiums, die Studie auf der Grundlage dieses Ergebnisses vorzeitig zu beenden, könnte zu einer falschen Schlussfolgerung über die Wirksamkeit der untersuchten Intervention führen.

Intuitiv betrachtet sollte die Durchführung zahlreicher Zwischenauswertungen bei $p<0{,}05$ den Fehler 1. Art (Rate



Im Laufe von 5 Jahren in Halbjahresintervallen durchgeführte Zwischenauswertungen. Angegeben ist der *p*-wert für den Vergleich zwischen der Behandlungs - und der Kontrollgruppe.

Tabelle 2. Nominelle Stoppgrenzen (*p*-Werte) für eine unterschiedliche Anzahl geplanter Zwischenauswertungen im Rahmen gruppensequenzieller Studienpläne [14,15].

Anzahl der geplanten Zwischenauswertungen	Zwischenauswertung	Pocock	Peto	O'Brien & Fleming
2	1	0,029	0,001	0,05
	2 (letzte)	0,029	0,005	0,048
3	1	0,022	0,001	0,0005
	2	0,022	0,001	0,014
	3 (letzte)	0,022	0,05	0,045
4	1	0,018	0,001	0,0001
	2	0,018	0,001	0,004
	3	0,018	0,001	0,019
	4 (letzte)	0,018	0,05	0,043
5	1	0,016	0,001	0,00001
	2	0,016	0,001	0,0013
	3	0,016	0,001	0,008
	4	0,016	0,001	0,023
	5 (letzte)	0,016	0,05	0,041

Gesamtirrtumswahrscheinlichkeit $\alpha = 0,05$

der falsch-positiven Ergebnisse), also α vergrößern. Tatsächlich aber steigt die tatsächliche Gesamt-Irrtumswahrscheinlichkeit α mit der Anzahl der Tests an, wenn ein Untersucher sich die bis dahin angesammelten Daten bei jeder Zwischenanalyse zum Niveau von $\alpha = 0,05$ ansieht, also beispielsweise: Gesamt-$\alpha = 0,08$ nach zwei Tests, $\alpha = 0,11$ nach drei und $\alpha = 0,19$ nach zehn Tests [9,13]. Dieses Multiplizitätsproblem verlangt nach statistischer Adjustierung: Davon hängt die wissenschaftliche Glaubwürdigkeit ab.

Methodiker haben zahlreiche statistische Stopp-Prozeduren (im Grunde eigentlich Warnregeln) entwickelt, die zuweilen auch datenabhängige Ablehnungsregeln oder Richtlinien genannt werden [13]. Wenn ein Untersucher Zwischenauswertungen vornimmt, muss er eine dieser Prozeduren anwenden. Gruppensequenzielle Studienpläne haben die vermutlich größte Aufmerksamkeit erlangt, denn meist sind sie recht einfach zu verstehen, aufzustellen und anzuwenden [14]. Ausgehend von der Anzahl der geplanten Zwischenauswertungen definieren diese Abbruchregeln, ab welchem *p*-Wert bei einer solchen Zwischenauswertung ein Studienabbruch in Erwägung gezogen werden sollte, während die Gesamt-Irrtumswahrscheinlichkeit 1. Art (α, Tabelle 2) erhalten bleibt.

Das Verfahren nach Pocock, das mit einem festgelegten nominellen Signifikanzniveau arbeitet, erweist sich als einfach und erlaubt die Beendigung einer Studie zu einem relativ frühen Zeitpunkt. Der Nachteil dieses Verfahrens besteht allerdings darin, dass der letzte Signifikanztest zu einem niedrigeren *p*-Wert durchgeführt werden muss als bei einer Studie mit einem gleichmäßigen, fixen Stichprobenumfang. Um beispielsweise bei drei Zwischenauswertungen eine Gesamt-Irrtumswahrscheinlichkeit von $\alpha = 0,05$ zu gewährleisten, müsste man bei jeder dieser Analysen, darunter auch der Endauswertung, zum Niveau von 0,022 testen (Tabelle 2). Ergäbe der letzte Test einen Wert von $p = 0,03$, würde man das Ergebnis der Studie nach diesem gruppensequenziellen Ansatz für nichtsignifikant erklären müssen. Das Ergebnis wäre jedoch signifikant gewesen, wenn man nicht nach einem gruppensequenziellen Plan vorgegangen wäre. Heute ist diese Methode hauptsächlich von historischem Interesse, denn mittlerweile haben sich andere Verfahren die Vorteile dieses Ansatzes zunutze gemacht, ohne dessen Nachteile in Kauf nehmen zu müssen [14].

Wir favorisieren zwei andere Verfahren, und zwar das von O'Brien & Fleming und Peto [12–14]. Beide wenden für die Zwischenauswertungen strenge Kriterien an (niedrige nominelle *p*-Werte) (Tabelle 2). Wird die Studie bis zum Erreichen des geplanten Stichprobenumfangs fortgeführt, dann werden alle Analysen so durchgeführt, als hätte es im Grunde keine Zwischenauswertungen gegeben. Beide Verfahren halten nicht nur das beabsichtigte Niveau ein, sondern auch die statistische Trennschärfe (Power) [16]. Die Daten werden im Wesentlichen genauso erhoben wie bei einem Studienplan mit fixem Stichprobenumfang. Ihr Reiz liegt in der Einfachheit. Die Implementierung einer Studie unter diesen Abbruchregeln spiegelt die einer herkömmlichen Studie wider, mit der einen Ausnahme, dass die Studie frühzeitig beendet wer-

den kann, falls sich eine Therapie als deutlich überlegen herausstellen sollte. Generell gilt, dass Untersucher kaum Vorteile erzielen, wenn sie während einer Studie mehr als vier oder fünf Zwischenauswertungen durchführen [9,17]. So können Untersucher mit nur geringem zusätzlichem Aufwand der ethischen Notwendigkeit einer Überprüfung der Studiendaten auf deutliche (positive oder negative) Therapieeffekte gerecht werden.

Leichter verstehen, implementieren und beschreiben lässt sich das Verfahren nach Peto (bzw. Haybittle & Peto). Es verwendet bis zur Schlussauswertung konstante, aber stringente nominelle Stoppgrenzen (Tabelle 2). In manchen Fällen kommen die Untersucher allerdings zu der Auffassung, dass die frühzeitige Beendigung einer Studie nach dem Peto-Verfahren zu schwierig ist.

Das Verfahren nach O'Brien & Fleming wirkt auf viele Untersucher intuitiv ansprechend, weil seine Abbruchkriterien im Frühstadium der Untersuchung, wenn man noch mit veränderlichen Ergebnissen rechnet, konservativ sind und nach und nach gelockert werden, wenn die Resultate zuverlässiger und stabiler werden. Die Stoppkriterien nach O'Brien & Fleming ändern sich im Gegensatz zum Peto-Verfahren mit jeder weiteren Zwischenauswertung.

Wenn Untersucher Interimsanalysen planen, sollten sie die statistischen Stoppregeln vorab festlegen. Außerdem sollte nicht der Untersucher, sondern ein unabhängiger Studienstatistiker die Auswertungen für das Datenüberwachungsgremium vornehmen [13]. Der Plan für die Zwischenauswertungen könnte ins Studienprotokoll, in einen separaten statistischen Analyseplan oder in die Satzung des Datenüberwachungsgremiums aufgenommen werden. Analyseplan und Satzung können dem Studienplan gegebenenfalls als Anhänge beigefügt werden. Auf diese Weise ist das Studienprotokoll für die Mitarbeiter, die die Studie durchführen, leichter zugänglich [13].

Bei der Mehrzahl der Studien sind wahrscheinlich keine Zwischenauswertungen und keine unabhängige Datenüberwachung erforderlich [18]. Von den 662 im Jahre 2000 identifizierten geeigneten Studien wurde in 24% der Fälle die Inanspruchnahme eines Datenüberwachungsgremiums bzw. die Durchführung von Zwischenauswertungen oder beides erwähnt [19].

Frühzeitige Beendigung der Studie und verzerrte Schätzer von Therapieeffekten

Beendet ein Datenüberwachungsgremium eine Studie auf der Basis einer gruppensequenziellen Stoppregel vorzeitig, sind die Effektschätzer verzerrt. Nach wie vor besteht darin eine Unzulänglichkeit eines solchen Vorgehens. Nehmen wir zur Erklärung einmal an, dass die Untersucher dieselbe Studie mehrmals durchgeführt hätten. Wahrscheinlich würden zufallsbedingte Schwankungen in Richtung eines größeren Behandlungseffektes eher zu einer frühzeitigen Beendigung der Studie führen als solche, die eine Tendenz zu schwächeren Therapieeffekten erkennen lassen.

Wenn eine Studie also vorzeitig beendet wurde, muss dem Leser klar sein, dass die geschätzten Therapieeffekte tendenziell überschätzt, d.h. zufallsbedingt hoch sind [12,14]. Wenn ein unverzerrter Effektschätzer sehr hoch ist, sollten die Untersucher sich auf einen Studienplan mit einem fixen Stichprobenumfang verständigen und gruppensequenzielle Designs vermeiden.

Beendigung der Studie wegen schädlicher Wirkungen oder Wirkungslosigkeit

Bislang sind wir in unserer Diskussion der Stoppregeln davon ausgegangen, dass unabhängig davon, ob ein Nutzen oder ein Schaden nachgewiesen wird, derselbe Grad der Evidenz nötig ist, um eine Studie vorzeitig abbrechen zu können. Methodiker bezeichnen eine solche Strategie bei gruppensequenziellen Verfahren als symmetrische Stoppgrenzen, in Analogie zum zweiseitigen Testen von Hypothesen.

Manche Untersucher oder Datenüberwachungsgremien bevorzugen mitunter aber asymmetrische Stoppgrenzen, die es ermöglichen, eine Studie beim Auftreten schädlicher Wirkungen mit einem niedrigeren Evidenzgrad vorzeitig zu beenden, als es beim Vorliegen positiver Wirkungen möglich wäre. So könnte man beispielsweise zur Überwachung von positiven Therapieeffekten die sequenziellen Annahmegrenzen der Regeln von O'Brien & Fleming anwenden, während beim Monitoring negativer Therapieeffekkte die Regeln nach Pocok angewendet werden könnten [13].

Manchmal wollen Wissenschaftler oder ein Datenüberwachungsgremium aber gar nicht schädliche Wirkungen nachweisen, sondern Trends aufzeigen, die so ungünstig sind, dass es unwahrscheinlich ist, dass man durch Weiterführung der Studie bis zu ihrem geplanten Ende einen signifikant nützlichen Effekt nachweisen kann. Das erleichtert den Abbruch einer Studie aufgrund von Wirkungslosigkeit, erlaubt aber nur die Feststellung, dass man nicht in der Lage war, eine positive Wirkung nachzuweisen. Das erzeugt so phantasievolle Wortschöpfungen wie „conditional power" (bedingte/konditionale Power) oder „stochastic curtailment" ("Schmälerung").

Beim Konzept der bedingten Power planen Untersucher ihre Studien mit einer festgelegten Trennschärfe [20]. Hat die Studie erst einmal begonnen, verbessert sich durch die fortwährende Akkumulation von Daten der Wissensstand (zunächst natürlich verborgen vor den Untersuchern). Mit zunehmend anfallenden Daten kann die statistische Power nun neu berechnet werden. Wenn sich beispielsweise ein Trend entwickelt, der die Wirksamkeit der Behandlung erkennen lässt, erhöht dies die Studienpower, während ein ungünstiger Trend sie verringert. Diese sich im Studienverlauf entwickelnde Abschätzung der Power wird durch das Konzept der bedingten Power beschrieben.

Überwachungsgremien verwenden dieses Konzept am häufigsten dann, wenn sich hinsichtlich der Behandlung ungünstige Trends abzeichnen. Wenn die Berechnungen der bedingten Power für verschiedene angenommene Behandlungseffekte (darunter auch für den im Studienprotokoll angenommenen The-

rapieeffekt) eine geringe Trennschärfe ergeben, dann könnte das Überwachungsgremium die Fortsetzung der Studie als nutzlos erachten und ihren Abbruch empfehlen. Eine solche Anwendung des Konzepts der bedingten Power hat Begriffe wie „stochastic curtailment" und „Abbruch wegen Wirkungslosigkeit" hervorgebracht. Diese Methoden sind bereits effektiv zur Überwachung von Studien eingesetzt worden [13,14].

Sonstige statistische Stoppregeln

Auch andere statistische Stoppregeln haben ihren Reiz. Lan & DeMets entwickelten mit der sog. „alpha spending function" eine Methode zur flexibleren Anpassung von gruppensequenziellen Methoden [21,22]. Diese Funktion kontrolliert die Irrtumswahrscheinlichkeit 1. Art (Anteil falsch-positiver Entscheidungen), die man bei jeder Zwischenauswertung annimmt, als Funktion des Anteils an der beobachteten Gesamtinformation. Mit dieser Methode können die Anzahl und die genauen Zeitpunkte der Zwischenauswertungen auch nach Studienbeginn noch verändert werden [13,14]. Das bedeutet, dass das Datenüberwachungsgremium mit einem Studienplan beginnt, der sich auf der Grundlage der im Studienverlauf erhobenen Daten noch ändern kann. Das Alpha-Spending-Verfahren ermöglicht also die ungeplante Ansicht von Daten.

Unserer Meinung nach eignen sich Bayesianische Ansätze für die klinische Entscheidungsfindung [23], wir sind aber nach wie vor skeptisch, was ihre Anwendung im Rahmen von Zwischenauswertungen betrifft. Bayes-Verfahren repräsentieren einen anderen Zweig der Statistik. Richtig implementiert können sie bei der Datenüberwachung von Nutzen sein [24–27]. Der Leser braucht sie aber nicht unbedingt zu verstehen, da sie nur selten zur Anwendung kommen. Überdies rufen sie Bedenken hervor. So könnte beispielsweise jede Zwischenauswertung auf dem Niveau von 0,05 durchgeführt werden, was die Gesamt-Irrtumswahrscheinlichkeit (Rate falsch-positiver Ergebnisse) stark ansteigen ließe [13]. Leider ist es aber so, dass manche Auftraggeber an dieser höheren Wahrscheinlichkeit, einen signifikanten Effekt zu finden, ein allzu starkes Interesse haben.

Worauf der Leser bei Zwischenauswertungen achten sollte

Der Leser sollte sich der Möglichkeit bewusst sein, dass Zwischenauswertungen nicht berichtet werden. Stößt er auf eine Aussage der Autoren, dass keine Zwischenauswertungen vorgenommen wurden, dann hat sich in der betreffenden Studie das Problem der Multiplizität wahrscheinlich gar nicht gestellt. Allerdings ist eine derart transparente Berichterstattung selten. Bei einer mangelhaften Berichterstattung dagegen könnte die Durchführung von Zwischenauswertungen verschleiert worden sein. Zugegeben, die Entdeckung solcher Zwischenauswertungen stellt für den Leser eine Herausforderung dar. Ein Anhaltspunkt könnte aber sein, dass der berechnete *p*-Wert nur geringfügig niedriger ist als 0,05, was bedeuten könnte, dass die Autoren wiederholt getestet und die Studie abgebrochen haben, nachdem ein *p*-Wert $<0{,}05$ erreicht worden war. Ein weiteres Indiz liegt möglicherweise vor, wenn die abschließende Studiengröße insgesamt kleiner war als geplant. Ein Grund, warum die Fallzahlberechnungen in den Methodenteil aufgenommen werden sollten, ist, dass sich daran erkennen lässt, ob die Studie zu einem vorzeitigen Ende gekommen ist. Der Leser sollte argwöhnisch werden, wenn die Studie vorzeitig beendet, aber keine statistische Stoppregel beschrieben wurde.

Wenn Untersucher eine statistische Stoppregel angeben, sollte der Leser nach der Angemessenheit eines solchen Vorgehens fragen. Mit den Verfahren nach Peto und O'Brien & Fleming lassen sich, ohne die Studie zu beeinträchtigen, dieselben Ziele erreichen wie mit Zwischenauswertungen. Die anderen statistischen Vorschläge zum Thema Zwischenauswertungen, von denen die meisten sich mit so phantasievollen Namen schmücken wie „alpha-spending function" und „conditional power", sind meist angemessen, die Bayesianischen Ansätze dagegen können Bedenken hervorrufen.

Erklärung zum Ausschluss von Interessenkonflikten

Hiermit erklären wir, dass keinerlei Interessenkonflikte vorliegen.

Danksagungen

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

Literatur

[1] Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. Lancet 2005;365:1591–5.

[2] Horton R. From star signs to trial guidelines. Lancet 2000;355:1033–4.

[3] Oei SG, Helmerhorst FM, Keirse MNC. Postcoital test should be performed as routine infertility test. BMJ 1999;318: 1008–1009ä.

[4] ISIS-2 Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. Lancet 1988;2:349–60.

[5] Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266: 93–8.

[6] Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002;21:2917–30.

[7] Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ 2003;326:219.

[8] Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000;355:1064–9.

[9] Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley; 1983.

[10] Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998;316:1236–8.

[11] Dyson DC, Crites YM, Ray DA, Armstrong MA. Prevention of preterm birth in high-risk patients: the role of education and provider contact versus home uterine monitoring. Am J Obstet Gynecol 1991;164:756–62.

[12] Pocock SJ. When to stop a clinical trial. BMJ 1992;305:235–40.

[13] Ellenberg SS, Fleming TR, DeMets DL. Data monitoring committees in clinical trials. Chichester: John Wiley and Sons; 2002.

[14] Piantadosi S. Clinical trials: a methodologic perspective. New York: John Wiley and Sons; 1997.
[15] Geller NL, Pocock SJ. Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. Biometrics 1987;43:213–23.
[16] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549–56.
[17] McPherson K. Sequential stopping rules in clinical trials. Stat Med 1990;9:595–600.
[18] Sydes MR, Spiegelhalter DJ, Altman DG, Babiker AB, Parmar MKB, and the DAMOCLES Group (2004) Systematic qualitative review of the literature on data monitoring committees for randomized controlled trials. *Clin Trials* 1: 60–79.
[19] Sydes MR, Altman DG, Babiker AB, Parmar MKB, Spiegelhalter D, and the DAMOCLES group (2004) Reported use of data monitoring committees in the main published reports of randomized trials: a cross-sectional study. *Clin Trials* 1: 48–59.
[20] Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005;365:1348–53.
[21] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659–63.
[22] DeMets DL, Lan KK. Interim analysis: the alpha spending function approach. Stat Med 1994;13:1341–52.
[23] Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. Lance 2005;365:1500–5.
[24] Spiegelhalter DJ, Freedman LS, Parmar MK. Applying Bayesian ideas in drug development and clinical trials. Stat Med 1993;12:1501–11.
[25] Freedman LS, Spiegelhalter DJ, Parmar MK. The what, why and how of Bayesian clinical trials monitoring. Stat Med 1994;13:1371–83.
[26] Parmar MK, Spiegelhalter DJ, Freedman LS. The CHART trials: Bayesian design and monitoring in practice. Stat Med 1994;13:1297–312.
[27] Parmar MK, Griffiths GO, Spiegelhalter DJ, Souhami RL, Altman DG, van der Scheuren E. Monitoring of large randomised clinical trials: a new approach with Bayesian methods. Lancet 2001;358:375–81.

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.

Literatur und Rezensionen

Reinhold Klein: 100 Fälle Allgemeinmedizin
1. Auflage, 432 Seiten, 100 Abb., Elsevier, München 2006. Preis: kartoniert € 24,95. ISBN 3-437- 43570 – 1

Dieses Buch entwickelt kasuistisch aus der hausärztlichen Praxis heraus die allgemeinmedizinische Berufstheorie nach R. N. Braun. Der Autor betrachtet seinen Praxisalltag mit einem gleichsam stereoskopischen Blick: einerseits hält er sich an den Braunschen Algorithmus, die Beratungsursachen werden zweiachsig systematisiert, nach der Häufigkeit ihres Auftretens („Senkrechte") und nach ihrer Klassifizierung durch den Allgemeinarzt in Symptom, Symptomgruppen, Bilder von Krankheiten und Diagnosen („Waagrechte"). Das ist sozusagen die makroskopische Oberfläche. Andrerseits liegt in der mikroskopischen Tiefe, der unmittelbaren Sinnlichkeit des „Falles", der in der Sprechstunde zur Vorstellung kommt, der zweite Schwerpunkt des Buches. Die Patienten treten uns sprachlich entgegen: z.B. appellativ „Herr Doktor, jetzt behandeln Sie erst meinen Rücken", aber auch relativ gefasst „Herr Doktor, mir brennt's beim Wasserlassen" oder auch indifferent „Herr Doktor – da haben Sie meinen Impfpass". Weiterhin im Photo als kindliches morbilliformes Amoxicillin-Exanthem z.B. als photographische Dokumentation eines depressiven Patienten im häuslichen „depressiven" Ambiente, auch dramatische Bilder von der Erstversorgung an Unfallorten, EKG-Befunde oder „roter Urin".

Klein nimmt sich der jeweiligen Patientenanliegen in der allgemeinärztlich typischen Manier an, die Anamnese ist fokussiert im Sinne einer verkürzten „Anamnestik", Befunde und deren Aussagekraft werden ausführlich diskutiert und in diagnostische und therapeutische Strategien umgesetzt. Wie das Vorgehen bei einem Patienten mit „Stechen in der Brust" belegen mag, geht es Klein auch um die Einarbeitung der kritischen Gewichtung von „Sensitivität" und „Spezifität" hausärztlicher diagnostischer Mittel – somit auch um evidence based Aspekte. Die besondere Spannung des Buchs liegt in der vielfältigen Verknüpfung einer allgemeinmedizinischen Berufstheorie mit der Praxis, aus der heraus sie gewonnen wurde und innerhalb derer sie täglich Anwendung findet. Immer wieder springt der Blick stereoskopisch zwischen der unmittelbaren Konkretion der vorgetragenen Anliegen, der Systematisierung dieser Anliegen als „Fälle" im Braunschen Algorithmus, der Benennung dieser hausärztlichen Fälle in der Nomenklatur der herkömmlichen Krankheitslehre. Klein verdeutlicht das durch ein dreifaches Inhaltsverzeichnis: Fall 1-100 („Herr Doktor…"), Übersicht über die Fälle nach der Braunschen „Kasugraphie", Übersicht über die Fälle nach Themen der herkömmlichen Krankheitslehre. Weiterhin verfügt das Buch über ein Glossar von Grundbegriffen in alphabetischer Folge, das dem ungeübten Leser Einführungen in die allgemeinmedizinische Berufstheorie nach Braun gibt und auch die aktuelle Verknüpfung zum ebenfalls epidemiologisch orientierten Paradigma der EBM herstellt. Viele Details eröffnen sich erst bei genauer Lektüre: die programmierte Diagnostik „Fieberstandard" nach Braun unter dem Gesichtspunkt unserer aktuellen Leitlinienkultur; Patienten, die mit verschiedenen Anliegen („Beratungsursachen") mehrfach auftauchen im Text. Schließlich die eindrucksvolle Interdisziplinarität des Autors (Durchsicht des Manuskripts durch spezialistische Kollegen, Beratung von Einzelthemen mit Vertretern anderer allgemeinmedizinischer Schulen), die auch für fachlich-wissenschaftliche Integrität steht. Ein Buch für den Medizinstudenten zur Lektüre oder Praktikumsbegleitung, für das Staatsexamen – hinsichtlich der Berufstheorie – genauso wie für den Weiterbildungsassistenten, Facharztprüfer Allgemeinmedizin und allgemeinmedizinische Hochschullehrer, zu denen Klein auch zählt. Letztere können sich hier einiges zu Didaktik und kasuistischer Aufbereitung von Prüfungsfragen an- wenn nicht gar abschauen.

Korrespondenzadresse:
PD Dr. Martin Konitzer
FA Allgemeinmedizin
Ferdinand-Wallbrecht-Str. 6-8
30163 Hannover
E-Mail: Nahid.Freudenberg@t-online.de