Vorsicht bei Subgruppenanalysen!

Peter Kleist
GlaxoSmithKline AG, Münchenbuchsee

Quintessenz

- In klinischen Studien können Subgruppenanalysen für ärztliche Therapieentscheidungen von Bedeutung sein, ihre Aussagekraft darf jedoch nicht überschätzt werden.
- Kleine Stichproben, ein Ungleichgewicht zwischen den Subgruppen und die steigende Anzahl statistischer Tests erhöhen die Wahrscheinlichkeit für Zufallsbefunde.
- Der Wert von Subgruppenanalysen basiert auf der Beachtung von Mindestanforderungen: Die Spezifizierung weniger Subgruppenanalysen im Studienprotokoll, die Durchführung von Interaktionstests und die statistische Korrektur für mehrfaches (multiples) Testen.
- Der Hauptzweck von Subgruppenanalysen in klinischen Studien besteht darin, Informationen über die Konsistenz eines Therapieeffektes zu erhalten. Unterschiede bezüglich Wirksamkeit oder Sicherheit einer Intervention in speziellen Subgruppen bieten meistens nur Anhaltspunkte, die mit Vorsicht zu interpretieren sind und der Bestätigung durch weitere Studien bedürfen.
- Das nachträgliche «Ausschlachten» von Daten mit der Suche nach Signifikanzen («data dredging»), manchmal mit dem Versuch, eine «negative Studie zu retten», entspricht keiner guten wissenschaftlichen Praxis.
- Die meisten Publikationen mit Subgruppenanalysen, selbst in guten medizinischen Fachzeitschriften, erfüllen die geforderten Standards nicht.

Summary

Subgroup analyses in clinical studies: handle with care

- *While analyses of treatment effects in subgroups of clinical studies may be of importance for therapeutic decision-making in individual patients, the importance of such analyses should not be overestimated.*
- *Small patient numbers, imbalances between subgroups and the growing range of statistical tests increase the probability of chance findings.*
- *The validity of subgroup analyses depends on compliance with basic requirements: prior specification of a small number of subgroup analyses in the protocol, conduct of treatment interaction tests and statistical adjustment for multiple testing.*
- *The main objective of subgroup analyses is to obtain information on the uniformity of the overall treatment effect. Subgroup differences in the efficacy or safety of an intervention are mainly of an exploratory character, should be interpreted with caution and need confirmation by subsequent clinical studies.*
- *Data dredging and fishing for statistical significances, sometimes for the purpose of "rescuing a negative study", are not sound scientific practices.*
- *Most publications containing subgroup analyses, even in high-ranking medical journals, fail to comply with current standards.*

Die Kontroverse

«Alle, die von diesem Heilmittel trinken, genesen innerhalb kurzer Zeit – diejenigen ausgenommen, bei denen es nicht hilft, und die alle sterben. Es ist daher offensichtlich, dass dieses Heilmittel nur in unheilbaren Fällen versagt.»
Galenos von Pergamon (Galen), 2. Jh. n. Chr.

«Subgroup analyses in clinical trials – fun to look at, but don't believe them.»
Professor Peter Sleight, Oxford, 20. Jh. n. Chr.

Ärzte und Statistiker haben zwar gelernt, miteinander zu leben, aber sie verhalten sich nicht immer wie Freunde. Ein Beispiel dafür ist die Kontroverse über den (Un-)Sinn einer Auswertung von Subgruppen in klinischen Studien, im angloamerikanischen Sprachraum auch als «clinicostatistical tragedy» bezeichnet [1]. Auf der einen Seite warnen Statistiker prinzipiell vor den Gefahren von Subgruppenanalysen: Ihr unverhältnismässig häufiger Gebrauch, auf reinem Zufall beruhende Befunde und/oder eine Überinterpretation der Ergebnisse können ungerechtfertigte Therapieempfehlungen zur Folge haben und dazu führen, bestimmten Patienten eine wirksame Therapie vorzuenthalten oder eine unwirksame Therapie zukommen zu lassen. Beispiele hierfür sind die auf Subgruppenanalysen basierenden vorübergehenden Beschränkungen der Thrombolyse und der Betablockergabe auf Patienten mit Vorderwandinfarkten oder einer Tamoxifentherapie bei Brustkrebs auf Patientinnen über 50 Jahre – Einschränkungen, die sich später als falsch erwiesen haben [2]. Die Schlussfolgerung der Kritiker von Subgruppenanalysen lautet daher: «Subgroups may kill people» [3].
Auf der anderen Seite steht der Arzt jedoch vor dem Problem, die Ergebnisse grosser randomisierter Studien mit oftmals sehr heterogenen Patientenkollektiven auf den klinischen Alltag und individuelle Patienten übertragen zu müssen. Denn die Nutzen-Risiko-Abwägung einer Therapie kann in Abhängigkeit vom Ausgangsrisiko oder vom Schweregrad einer Erkrankung unterschiedlich ausfallen. Beispiele hierfür sind die Antikoagulation bei unkompliziertem bzw. risikobehaftetem Vorhofflimmern oder die Karotisthrombendarterektomie bei unterschiedlichen

CME zu diesem Artikel finden Sie auf S. 791 oder im Internet unter www.smf-cme.ch.

Stenosierungsgraden. Ist davon auszugehen, dass eine Therapie für bestimmte Patientengruppen mehr oder weniger wirksam oder sogar gefährlich sein kann, dann besteht eine wissenschaftliche und ethische Verpflichtung, solche Subgruppen zu identifizieren [4]. Daten zu spezifischen Subgruppen von Patienten sind somit ein wichtiger Bestandteil ärztlicher Therapieentscheidungen.

Sind beide Sichtweisen tatsächlich unvereinbar? Der folgende Beitrag geht davon aus, dass sinnvolle Subgruppenfragestellungen ein essentieller Bestandteil grosser randomisierter Studien sind. Klinische Bedürfnisse und methodologische Anforderungen müssen jedoch miteinander verbunden werden, um Subgruppenergebnissen letztlich eine wissenschaftliche Aussagekraft und somit überhaupt eine klinische Bedeutung zukommen zu lassen. Nach einigen «warnenden Beispielen» zu Beginn sollen Möglichkeiten und Grenzen bei der Durchführung und Interpretation von Subgruppenanalysen aufgezeigt werden.

Probleme mit Subgruppenanalysen

«Der Zufall arbeitet viel verlässlicher, wenn man ihm auf die Sprünge hilft.»

Henryk Bereska (1926–2005), Lyriker und Übersetzer polnischer Literatur

Das konventionelle Signifikanzniveau eines statistischen Tests von 5% besagt, dass mit einer Wahrscheinlichkeit von 5% mit einem zufälligen, falschpositiven Ergebnis zu rechnen ist. Auf den ersten Blick scheint dies wenig zu sein. Aber Vorsicht: Die Wahrscheinlichkeit hierfür ist nahezu doppelt so hoch wie jene, mit zwei Würfeln in einem Wurf zwei «Sechsen» zu erhalten. Werden in einer klinischen Studie – zusätzlich zur primären Fragestellung – noch verschiedene Subgruppen ausgewertet, ist dies mit mehrmaligem Würfeln vergleichbar: Die Anzahl statistischer Tests steigt und mit ihr die Zufallswahrscheinlichkeit. Bei fünf Subgruppenanalysen beträgt die Wahrscheinlichkeit für *ein* zufälliges (falschpositives) Ergebnis bereits 23%, bei zehn Auswertungen sind es 40, bei 20 Auswertungen 64 und bei 50 Auswertungen sogar 92%. Es verwundert daher nicht, dass die Betablocker-Heart-Attack-Studie [5], in der 146 Subgruppen ausgewertet wurden, auch einige irreführende Ergebnisse produzierte. Hinzu kommt, dass der Stichprobenumfang in Subgruppen relativ klein ist und die Teststärke (Power) für Subgruppenanalysen rapide abnimmt. Die Wahrscheinlichkeit, einen tatsächlich bestehenden Unterschied zwischen Subgruppen aufdecken zu können, ist also gering.

Ist eine Behandlung wirksam und das Gesamtergebnis einer klinischen Studie signifikant, muss man zwangsläufig mit Subgruppen rechnen, bei denen die Behandlung scheinbar nicht wirkt (z.B. Betablocker bei Hinterwandinfarkten) – vor allem dann, wenn diese Subgruppen klein sind und aus unterrepräsentierten Patientengruppen wie Frauen oder älteren Menschen bestehen. Gemäss statistischer Simulationen beträgt die Wahrscheinlichkeit für eine Subgruppe mit nichtsignifikantem Studienausgang – und das bei nur einer Subgruppenfragestellung – bis zu 66%. Fällt das Gesamtergebnis einer Studie negativ aus, beträgt die Wahrscheinlichkeit, eine Subgruppe mit einem signifikanten Ergebnis zu erhalten, immer noch bis zu 21% [6].

Wie der Zufall am Werk ist, soll mit den folgenden Beispielen erläutert werden. Die ISIS-2-Studie [7] an über 17 000 Patienten wies eindrucksvoll nach, dass Acetylsalicylsäure (ASS) die Mortalität nach einem Myokardinfarkt senkt. Gemäss einer Subgruppenanalyse auf der Basis des Sternzeichens der Studienteilnehmer scheint ASS bei Patienten mit den Sternzeichen Waage oder Zwillinge jedoch nicht wirksam zu sein (Tab. 1) [8]. Ein ähnliches Beispiel betrifft die Wirksamkeit der Karotisthrombendarterektomie zur Hirnschlagprophylaxe bei hochgradigen symptomatischen Stenosen: Bei Patienten, die in den Monaten Mai und Juni geboren wurden, scheint das chirurgische Verfahren besonders effektiv zu sein, während es für im März und April geborene Patienten potentiell gefährlich ist (Abb. 1) [2, 9]. Sie mögen lachen, denn niemand (ausser vielleicht ein Astrologe) käme auf die Idee, dass der klinische Nutzen der beiden genannten Verfahren vom Sternzeichen bzw. vom Geburtsmonat der Patienten abhängig ist. Erschiene dagegen ein Zusammenhang plausibel, zum Beispiel in bezug auf das Lebensalter oder den Blutdruck, wären wir jedoch viel schneller geneigt, solche Ergebnisse für klinisch bedeutsam zu erachten.

Tabelle 1. Einfluss von Acetylsalicylsäure (ASS) auf die Mortalität nach einem Myokardinfarkt (Subgruppenauswertung der ISIS-2-Studie auf Basis des Sternzeichens der Studienteilnehmer [7, 8]).

Sternzeichen	Vaskuläre Todesfälle nach einem Monat		Signifikanzniveau
	ASS	Plazebo	
Waage oder Zwillinge	150 (11,1%)	147 (10,2%)	p = 0,5
Alle anderen Sternzeichen	654 (9,0%)	869 (12,1%)	p <0,0001
Alle Patienten (alle Sternzeichen)	804 (9,4%)	1016 (11,8%)	p <0,0001

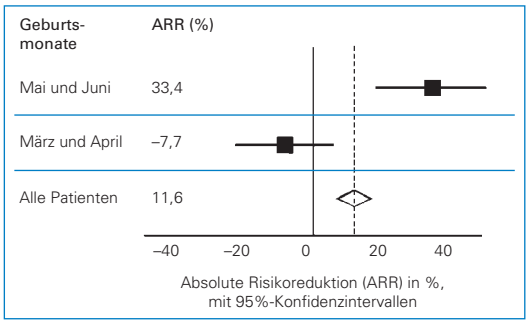

Abbildung 1
Effekte der Karotisthrombendarterektomie bei Patienten mit einer ≥70%igen symptomatischen Stenose in der ECS-Studie [9]. Darstellung ausgewählter Ergebnisse einer Subgruppenauswertung auf Basis des Geburtsmonats der Patienten (adaptiert nach [2]).

Mitunter braucht es nicht einmal medizinische Plausibilität, um die Ergebnisse von Subgruppenanalysen über Gebühr ernst zu nehmen. In der PRAISE-1-Studie [10] wurde der Kalziumantagonist Amlodipin an über 1100 herzinsuffizienten Patienten mit Plazebo verglichen. In Bezug auf den primären Endpunkt (Kombination aus Gesamtsterblichkeit und kardiovaskulärer Morbidität) war kein Unterschied zwischen Amlodipin und Plazebo zu sehen. Da die Studienärzte wohl zuvor davon ausgegangen waren, dass Patienten mit einer ischämisch bedingten Herzinsuffizienz eher von der Therapie profitieren könnten, sah das Protokoll einen stratifizierten Patienteneinschluss und eine entsprechende Subgruppenauswertung nach der Ursache der Herzinsuffizienz vor (ischämisch vs. nichtischämisch). Unerwarteterweise war in der Patientengruppe mit einer nichtischämischen dilatativen Kardiomyopathie durch Amlodipin im Vergleich zu Plazebo eine Risikoreduktion von 31% nachweisbar, während sich in Bezug auf die Patienten mit ischämisch bedingter Herzinsuffizienz kein Unterschied zeigte. Zur «Bestätigung» dieses Subgruppeneffektes wurde eine Folgeuntersuchung, und zwar die PRAISE-2-Studie [11], bei herzinsuffizienten Patienten und einem normalen Koronarangiogramm durchgeführt. Das Ergebnis dieser zweiten Studie konnte das Ergebnis der ersten nicht bestätigen – im Gegenteil, es war sogar ein Trend in Richtung einer erhöhten Mortalität unter dem Kalziumantagonisten zu verzeichnen (Tab. 2). Das Subgruppenergebnis von PRAISE-1 beruhte also auf reinem Zufall.

Werden Subgruppenanalysen erst im nachhinein, das heisst nach der Kenntnis der Datenlage vorgenommen (sog. Post-hoc-Analysen), ist die Wahrscheinlichkeit für Zufallsergebnisse noch grösser. Wir haben dies bereits bei den nachträglichen Auswertungen zum Sternzeichen und zum Geburtsmonat gesehen. Wie gefährlich Post-hoc-Analysen sein können, verdeutlicht das Beispiel der Canadian Cooperative Study zu Acetylsalicylsäure bei Patienten mit transitorisch ischämischen Attacken, einer in den 1970er Jahren richtungsweisenden Studie [12]: Da der Effekt von ASS nur bei Männern statistische Signifikanz erreichte, war die Substanz in den USA für lange Zeit (bis weitere Studien das Gegenteil bewiesen) zur Schlaganfallprophylaxe nur für Männer zugelassen – was dazu führte, dass vielen Frauen eine potentiell lebensrettende Therapie vorenthalten wurde [8].

Anforderungen an Subgruppenanalysen

Die prinzipiellen Anforderungen an Subgruppenanalysen lassen sich aus den oben beschriebenen Beispielen ableiten und sind in Tabelle 3 übersichtlich aufgeführt. An dieser Stelle sollen nur einige Erläuterungen gegeben werden. Allen Anforderungen gemein ist eine (weitgehend) effektive Kontrolle des Zufalls, damit Subgruppenanalysen ein akzeptables Mass an Verlässlichkeit und Glaubwürdigkeit zukommt. Dies bedeutet in der Regel, sich auf wenige, klinisch relevante Fragestellungen mit einer klaren biologischen Rationale zu beschränken [2, 13] – was so manchem Studienleiter schwerfällt, da in grossen Studien eine immense Menge an Patientendaten erhoben wird. Aber hier bestätigt sich wieder einmal, dass «weniger oft mehr ist». Die Hinzunahme von weniger relevanten Fragestellungen erhöht die Anzahl statistischer Tests und damit die Zufallswahrscheinlichkeit, was die Aussagekraft der wirklich wichtigen Fragestellungen ungewollt einschränkt.

Tabelle 2. Amlodipin vs. Plazebo bei Patienten mit Herzinsuffizienz (HI). Ergebnisse einer Subgruppenanalyse bei Patienten ohne ischämische Herzerkrankung PRAISE-1 [10] und der daraufhin durchgeführten Studie PRAISE-2 [11] an Patienten mit Herzinsuffizienz und einem normalen Koronarangiogramm.

	PRAISE-1		PRAISE-2		PRAISE-1 + PRAISE-2 kombiniert	
	Subgruppe Nur Patienten mit nichtischämischer HI		*Alle Patienten der Studie* Normales Koronarangiogramm			
Therapie	Plazebo	Amlodipin	Plazebo	Amlodipin	Plazebo	Amlodipin
Patientenanzahl	212	209	826	826	1408	1397
Mortalität	34,9%	21,5%	31,7%	33,7%	34,0%	33,4%

Tabelle 3. Zehn (Ideal-)Anforderungen an die korrekte Durchführung von Subgruppenanalysen. Die (aus der Sicht des Autors) Wichtigsten sind fettgedruckt.

1.	**Festlegung geplanter Subgruppenanalysen im Studienprotokoll (oder in einem separaten Analyseplan);** Formulierung von Hypothesen und Beschreibung der Effekte (Richtung und Ausmass)
2.	Darlegung der Rationale für die Auswertung von Subgruppen (erwartete biologische Wirkung)
3.	**Exakte Definition der einzelnen Subgruppen im voraus** (z.B. Quartile oder Unter-/Obergrenzen für Messwerte)
4.	**Beschränkung der Analyse auf wenige, relevante Subgruppen** (Subgruppeneffekte sollten nur in Bezug auf den primären Studienendpunkt untersucht werden)
5.	Berücksichtigung bei der Fallzahlplanung, falls wesentliche Subgruppeneffekte erwartet werden (zur Gewährleistung einer adäquaten Teststärke)
6.	Eine stratifizierte Randomisierung sollte in Betracht gezogen werden
7.	**Beschränkung der Analyse von Subgruppen auf Baseline-Variablen**
8.	**Durchführung eines Interaktionstests als erster Auswertungsschritt** (erst ein signifikanter Interaktionstest erlaubt eine sich anschliessende, separate Auswertung von Subgruppen)
9.	**Statistische Korrektur für mehrfaches (multiples) Testen**
10.	Erwähnung aller durchgeführter Subgruppenanalysen in der Publikation

Die Richtlinien der Arzneimittelbehörden [14, 15] fordern, dass Subgruppenanalysen im voraus zu planen und bereits im Studienprotokoll (oder in einem Analyseplan), zumindest vor der Kenntnis der Studiendaten, festzulegen sind. Datengetriebene Post-hoc-Analysen werden von den Behörden nicht akzeptiert, da die Wahrscheinlichkeit für falschpositive Ergebnisse besonders hoch ist. Es lässt sich zudem nicht überprüfen, wie viele Tests tatsächlich durchgeführt wurden und wie viele nichtsignifikante Ergebnisse unter den Tisch gefallen sind. Auch ausserhalb des regulatorischen Bereichs werden Post-hoc-Auswertungen kritisch beurteilt [8, 16, 17]. Das CONSORT Statement [18], auf das sich die führenden biomedizinischen Zeitschriften bei der Bewertung von Publikationsmanuskripten stützen, erfordert eine klare Angabe darüber, welche Analysen präspezifiziert waren und welche nachträglich vorgenommen wurden.

Der erste Schritt einer Subgruppenauswertung muss grundsätzlich in der Durchführung eines Interaktionstests (Test auf Heterogenität des Behandlungseffektes) bestehen. Erst wenn ein Interaktionstest signifikant ist, können anschliessend separate Subgruppen analysiert werden. Der primäre Ansatz besteht also darin, das Ausmass der Behandlungseffekte miteinander zu vergleichen, das heisst die relativen Risiken (RR) oder odds ratios (OR) – und nicht die p-Werte von Subgruppen [19]. Was bedeutet das jetzt konkret? Ein hypothetisches Beispiel soll dies erläutern: Angenommen, eine neue Osteoporosebehandlung reduziert das Auftreten einer nichtvertebralen Fraktur gegenüber der Kontrollbehandlung signifikant um 20% (RR: 0,8; CI: 0,71–0,89; p <0,05). Bei Frauen mit einer vorbestehenden Fraktur (Subgruppe 1) beträgt das relative Risiko 0,77 (0,57–0,97; p <0,05), bei Frauen ohne vorbestehende Fraktur (Subgruppe 2) hingegen «nur» 0,83 (0,59–1,07; nichtsignifikant). Heisst dies, dass die neue Behandlung nur in Subgruppe 1, nicht jedoch in Subgruppe 2 wirksam ist? Die relevante Frage ist, ob sich die Risikoreduktion von 23% in Subgruppe 1 von derjenigen von 17% in Subgruppe 2 unterscheidet. Der Interaktionstest, der die beiden relativen Risiken miteinander vergleicht, ist jedoch nicht signifikant, und die zuvor gegenübergestellten ermittelten p-Werte der separaten Subgruppen (signifikant in Subgruppe 1 und nichtsignifikant in Subgruppe 2) dürfen nicht zur Schlussfolgerung verleiten, dass sich die Behandlungseffekte voneinander unterscheiden [20]. Solche Fehler finden sich jedoch häufig in der Literatur. Beispielsweise wurde in der Publikation einer Studie an Infarktpatienten zu unrecht behauptet, dass die zusätzliche psychologische Betreuung mit einem erhöhten Risiko für Frauen verbunden sei [21]. Zwar erreichte eine scheinbar erhöhte kardiale Mortalität nur bei Frauen statistische Signifikanz, ein (leider nicht durchgeführter) Interaktionstest hätte allerdings gezeigt, dass sich die Behandlungseffekte bei Frauen und Männern nicht signifikant voneinander unterscheiden [4].

Interaktionstests dämmen zwar die «Flut» zufälliger Ergebnisse ein, stellen aber letztlich keine Garantie dar. Denn auch ein signifikanter Interaktionstest schützt nicht vor falschen Schlussfolgerungen; im Beispiel der Abbildung 1 ist der Interaktionstest mit p <0,0001 hochsignifikant. Zudem ist die Teststärke (Power) von Interaktionstests gering, da eine klinische Studie in aller Regel nur für die primäre Fragestellung «gepowert» wurde; sie nimmt von beispielsweise 80% für den Behandlungseffekt in der Gesamtstudie auf maximal 29% für einen Interaktionseffekt mit dem gleichen Ausmass ab [22]. Die andererseits zur Beibehaltung der Teststärke notwendige vierfache Erhöhung der Fallzahl in der Studie ist jedoch als eine unrealistische Alternative zu betrachten.

Mehrfaches Testen im Rahmen von Subgruppenauswertungen erhöht die Wahrscheinlichkeit für falschpositive Resultate und erfordert eine statistische Korrektur. Erfolgen keine hierarchischen Testprozeduren (d.h. der nächste Test darf nur durchgeführt werden, wenn der vorangehende signifikant war), ist das Signifikanzniveau entsprechend der Anzahl der Tests anzupassen. Tabelle 4 gibt einen Überblick darüber, wie sich das tatsächliche Signifikanzniveau eines Tests in Abhängigkeit von der Anzahl der analysierten Subgruppen verändert und welche Anpassungen notwendig sind, um ein nominelles Signifikanzniveau für jeden Einzeltest beizubehalten. Dass solche Korrekturen auch für die Durchführung mehrerer Interaktionstests erforderlich sind, zeigt das folgende Beispiel: In einer Vergleichsstudie zwischen der Kombination von Clopidogrel und Acetylsalicylsäure und einer ASS-Monotherapie konnte kein signifikanter Unterschied zwischen

Tabelle 4. Signifikanzniveau in Abhängigkeit von der Anzahl untersuchter Subgruppen.

Anzahl der Subgruppen	Tatsächliches Signifikanzniveau ohne statistische Korrektur	Signifikanzniveau des Einzeltests zur Beibehaltung eines Gesamtniveaus von 0,05
2	0,0975	0,0253
3	0,1426	0,0170
4	0,1855	0,0127
5	0,2262	0,0102
6	0,2649	0,0085
7	0,3017	0,0073
8	0,3366	0,0064
9	0,3598	0,0057
10	0,4013	0,0051

Nominelles Signifikanzniveau = 0,05; mittlere Spalte: tatsächliches Signifikanzniveau ohne Korrektur für multiples Testen; rechte Spalte: angepasstes Signifikanzniveau des einzelnen Tests, um insgesamt ein Signifikanzniveau von 0,05 zu bewahren.

beiden Behandlungen bezüglich der Verhinderung atherothrombotischer Ereignisse festgestellt werden [23]. Nur eine von 20 Subgruppenanalysen (symptomatische Patienten) ergab überhaupt einen statistisch signifikanten Vorteil für eine der Behandlungen (Interaktionstest mit p = 0,045). Hätte man allerdings in angemessener Weise das Interaktionssignifikanzniveau für 20 durchgeführte Tests korrigiert (0,05/20 = 0,0025), hätte kein Subgruppenergebnis auch nur annähernd statistische Signifikanz erreicht [24]. Subgruppenanalysen müssen sich auf Patienteneigenschaften beschränken, die bereits vor Beginn der Studie vorlagen (z.B. Alter, Erkrankungsschweregrad oder Risikofaktoren). Es dürfen keine Variablen sein, die potentiell durch die Behandlung beeinflusst werden und darüber entscheiden können, ob ein Patient einer bestimmten Subgruppe zugehört oder nicht (Beispiele hierfür wären veränderte Laborvariablen oder klinische Symptome) [25]. Vergleiche auf der Basis von Post-Randomisierungsbefunden unterliegen potentiell einem Bias, beispielsweise wenn in einer plazebokontrollierten Studie zu einem Statin Therapieresponder nach dem Ausmass der erzielten Cholesterinsenkung definiert werden [26]. Das Auftreten eines klinischen Endpunktes (z.B. ein kardiovaskuläres Ereignis) kann eng mit der biochemischen Response verbunden sein und stellt daher keine unabhängige Untersuchungsvariable mehr dar. Auswertungen von Responder- und Nonrespondersubgruppen sind somit nicht valide. Ähnlich problematisch sind auch Subgruppenauswertungen zur Therapie-Compliance, da die Patienten der Behandlungsgruppe aufgrund unerwünschter Wirkungen und jene der Plazebogruppe hingegen aufgrund mangelnder Wirksamkeit, das heisst also aus unterschiedlichen Gründen, eine schlechte Compliance aufweisen oder sogar vorzeitig aus der Studie ausscheiden können. Die einander gegenübergestellten Subgruppen sind daher nicht mehr vergleichbar [8]. Beispielhaft hierfür ist eine im Coronary Drug Project vorgenommene Auswertung [27]: Patienten in der Clofibratgruppe, die mindestens 80% der Studienmedikation einnahmen, hatten eine hochsignifikant tiefere Fünf-Jahres-Mortalität als Patienten mit schlechter Compliance (15 vs. 24,6%; p = 0,0001). Allerdings war der Unterschied bei den Patienten in der Plazebogruppe noch eindrücklicher (15,1 vs. 28,3%; p <0,00001) – scheinbar haben Patienten mit einer regelmässigen Plazeboeinnahme den grössten Nutzen.

Zusammenfassende Bewertung und Empfehlungen

«The best test of the validity of subgroup analyses is not significance but replication.»
Professor Peter M. Rothwell, Oxford

Systematische Untersuchungen zur Qualität von Subgruppenanalysen bestätigen, dass die meisten veröffentlichten Studien – selbst in hochrangigen medizinischen Fachzeitschriften – den geforderten Standards nicht genügen [4, 13, 22, 28]. Von Subgruppenanalysen wird übermässig Gebrauch gemacht und die Ergebnisse aufgrund methodischer Mängel häufig überbewertet. Ein Interaktionstest wird oftmals nicht durchgeführt, so dass sich die Darstellung der Ergebnisse auf den Vergleich von (unadjustierten) p-Werten einzelner Subgruppen beschränkt. Die Wahrscheinlichkeit für Zufallsergebnisse und ungerechtfertigte Therapieempfehlungen ist somit sehr gross. Besondere Skepsis ist bei Post-hoc-Analysen angebracht, insbesondere dann, wenn die Ergebnisse der Subgruppen dem in der Studie beobachteten Gesamteffekt widersprechen und mit einer Subruppenauswertung der Versuch unternommen wird, eine «negative Studie zu retten». Das nachträgliche «Ausschlachten von Daten» mit der Suche nach statistischen Signifikanzen (sog. «data dredging») kommt einem Missbrauch von Subgruppenanalysen gleich und ist als unwissenschaftlich zu taxieren [16].

Beobachtete Signifikanzen im Interaktionstest beruhen meistens nur auf quantitativen Interaktionen, das heisst dass das Ausmass des Behandlungseffektes in einer Subgruppe grösser ist als in einer anderen. Diese Interaktionen sind häufig zufallsbedingt und in der Regel klinisch irrelevant. Sogenannte qualitative Interaktionen, das heisst dass die eine Behandlung in der einen Subgruppe eindeutig besser ist und in der anderen eindeutig unterlegen (oder sogar gefährlich), sind dagegen äusserst selten [28].

Bei kleinen Studien ist von Subgruppenanalysen generell abzusehen. Subgruppenanalysen innerhalb von grossen randomisierten und multizentrischen Studien können nur dann klinisch bedeutsam sein, wenn die dargelegten Durchführungsanforderungen Beachtung finden. Ist eine Subgruppe von sehr hoher klinischer Relevanz, sollte die Randomisierung der Patienten vorzugsweise stratifiziert erfolgen, um einem möglichen Randomisierungsbias zu begegnen. Die Bedeutung von Subgruppenauswertungen darf aber auch dann nicht überschätzt werden, weil der beobachtete Gesamteffekt in der Studie aufgrund der höheren Power immer noch zuverlässiger ist als jener in den einzelnen Subgruppen [8].

Die grösste Aussagekraft kommt Subgruppenanalysen bei der Frage zu, ob sie den Gesamteffekt einer Studie und dessen Konsistenz in verschiedenen Patientenkollektiven grundsätzlich unterstützen. Unterschiede bezüglich der Wirksamkeit oder Sicherheit in wenigen, speziellen Patientengruppen haben nur explorativen Charakter und sind mit grosser Vorsicht zu interpretieren. Das Vorhandensein einer klaren biologischen Rationale, Hinweise auf eine Dosis-Wirkungs-Beziehung oder die Reproduzierbarkeit der Beobachtung durch andere Stichproben, zum Beispiel durch eine Auswertung auf der Basis der einzelnen Studienzentren, können die Validität eines Subgruppenergebnisses bedingt stützen [25]. Solche Anhaltspunkte für Subgruppeneffekte dienen aber letztlich nur der Hypothesengenerierung und sind am besten durch eine weitere, unabhängige Studie zu überprüfen [2, 22]. Auch die Durchführung einer Metaanalyse unter Berücksichtigung der Subgruppen verschiedener Studien kann ein sinnvoller Ansatz sein. Dass der Nutzen der Thrombolyse bei einem Myokardinfarkt umso grösser ist, je früher die Behandlung einsetzt – um ein Beispiel zu nennen –, konnte erst durch eine Metaanalyse grosser Studien demonstriert werden [29].

Korrespondenz:
Dr. med. Peter Kleist
Viktoriastrasse 52
CH-3084 Wabern
peter.kleist@bluewin.ch

Empfohlene Literatur

– Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity. I: clinical trials. Lancet. 2001;357:373–80.
– Counsell CE, Clarke MJ, Slattery J, Sandercock PAG. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? BMJ. 1994;309:1677–81.

– Das vollständige Literaturverzeichnis [1–29] finden Sie in der Onlineausgabe dieses Artikels unter www.medicalforum.ch/pdf/pdf_d/2007-39-210.pdf.

Vorsicht bei Subgruppenanalysen in klinischen Studien

Peter Kleist

GlaxoSmithKline AG, Münchenbuchsee

Korrespondenz:
Dr. Peter Kleist MD, FFPM
Medical Director
GlaxoSmithKline AG
Talstrasse 3-5
CH-3053 Münchenbuchsee
peter.m.kleist@gsk.com

Literatur

1 Feinstein AR. The problem of cogent subgroups: A clinicostatistical tragedy. J Clin Epidemiol 1998;51:297–9.
Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet. 2005;365:176–86.
2 Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretations. Lancet 2005;365:176–86.
3 van Gijn J. Extrapolation of trial data into practice: where is the limit? Cerebrovasc Dis 1995;5:159–62.
4 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.
5 Betablocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. JAMA 1982; 247: 1707-14.
6 Brookes ST, Whitley E, Peters TJ, et al. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess 2001;5:1–56.
7 ISIS-2 (Second International Study on Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2. Lancet. 1988;2:349–60.
8 Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. Lancet. 2001;357:373–80.
9 European Carotid Surgery Trialists' Collaborative Group. Randomised trial of endarterectomy for recently symptomatic carotid stenosis: final results of the MRC European Carotid Surgery Trial (ECST). Lancet. 1998;351:1379–87.
10 Packer M, O'Connor CM, Ghali JK, et al. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. N Engl J Med. 1996;335:1107–14.
11 Thackray S, Witte K, Clark AL, Cleland JGF. Clinical trials update: OPTIME-CHF, PRAISE-2, ALL-HAT. Eur J Heart Fail. 2000;2:209–12.
12 The Canadian Cooperative Study Group. A randomised trial of aspirin and sulfinpyrazone in threatened stroke. N Engl J Med. 1978;299:53–9.
13 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.
14 International Conference on Harmonisation (ICH). Guideline E9: Statistical principals for clinical trials. 05 February 1998. www.ich.org/LOB/media/MEDIA485.pdf (Zugriff am 14. Juni 2007).
15 Committee for Proprietary Medicinal Products (CPMP). Points to consider on multiplicity issues in clinical trials. CPMP/EWP/908/99. 19 December 2002. www.emea.europa.eu/pdfs/human/ewp/090899en.pdf (Zugriff am 14. Juni 2007).
16 Lewis SC, Warlow CP. How to spot bias and other potential problems in randomised controlled trials. J Neurol Neurosurg Psychiatry. 2004;75:181–7.
17 Sleight P. Debate: Subgroup analyses in clinical trials – fun to look at, but don't believe them. Curr Control Cardiovasc Med. 2000;1:25–7.
18 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.
19 Matthews JNS, Altman DG. Interaction 2: compare effect sizes not P values. BMJ. 1996;313:808.
20 Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ. 2003;326:219.
21 Frasure-Smith N, Lespérance F, Prince RH, et al. Randomised trial of home-based psychological nursing intervention for patients recovering from myocardial infarction. Lancet. 1997;350:473–9.
22 Hernandez AV, Boersma E, Murray GD, et al. Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? Am Heart J. 2006;151:257–64.
23 Bhatt DL, Fox KAA, Hacke W, et al. Clopidogrel and aspirin versus aspirin alone for the prevention of atherothrombotic events. N Engl J Med. 2006;354:1706–17.
24 Lagakos SW. The challenge of subgroup analyses – reporting without distorting. N Engl J Med 2006;354:166–9.
25 Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. MJA 2004;180:289–91.
26 West of Scotland Coronary Prevention Study Group. Influence of pravastatin and plasma lipids on clinical events in the West of Scotland Coronary Prevention Study (WOSCOPS). Circulation 1998;97:1440–5.
27 The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. N Engl J Med. 1980;303:1038–41.

28 Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: Some lessons from major cardiovascular trials. Am Heart J 2000; 139:952–61.

29 Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. Indications for fibrinolytic therapy in suspected myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. Lancet. 1994;343:311–22.