Bundesgesundheitsbl 2015 · 58:274–282
DOI 10.1007/s00103-014-2105-2
Online publiziert: 8. Januar 2015

A. Gonnermann · M. Kottas · Armin Koch
Institut für Biometrie, Medizinische Hochschule Hannover, Hannover, Deutschland

Biometrische Entscheidungsunterstützung in Zulassung und Nutzenbewertung am Beispiel der Implikationen von heterogenen Ergebnissen in Untergruppen der Studienpopulation

Auch um dem Vorwurf zu begegnen, dass in einer randomisierten klinischen Studie neue Arzneimittel in einem „Windkanalversuch" untersucht werden, ist es Paradigma der modernen Therapieforschung, ein neues Arzneimittel in Phase III in einer möglichst breiten Population zu untersuchen, die nur durch wenige Ein- und Ausschlusskriterien eingeschränkt wird. Diese Ein- und Ausschlusskriterien werden üblicherweise z. B. durch mechanistische Überlegungen zur Wirksamkeit des Arzneimittels oder durch Sicherheitsüberlegungen begründet. Ärzte sind darauf trainiert, eine optimale Therapie unter Berücksichtigung der spezifischen Indikationen und Kontraindikationen eines Patienten zu wählen und werden sich folglich für Untergruppen der Patientenpopulation interessieren, die sie als klinisch oder prognostisch relevante Entitäten wahrnehmen. Alter, Geschlecht, Tumorstatus und Karnofsky-Index sind nur Beispiele für Faktoren, die prognostisch für den Verlauf vieler Erkrankungen sind.

Hoffnung und Ausgangspunkt vieler Arzneimittelentwicklungen ist die Annahme, dass diese verschiedenen prognostisch relevanten Untergruppen konsistent von einem Arzneimittel oder einer Therapie profitieren *und* in diesen Untergruppen das Nutzen-Risiko-Verhältnis positiv ist. In der Planungsphase einer klinischen Studie wird man also versuchen, die möglichst größte Population zu identifizieren, für die die vorgenannten Kriterien erfüllt sind. Es ist jedoch quasi der Preis der Entwicklung eines Arzneimittels für eine möglichst uneingeschränkte Population, dass die Annahme der Konsistenz oder Uniformität des Behandlungseffekts und die gleichmäßig sichere Anwendung des Arzneimittels post hoc zu überprüfen sind.

Eine ähnliche Situation findet sich in der Bewertung des Zusatznutzens von Arzneimitteln, die gegebenenfalls zur Folge hat, dass höhere Kosten gerechtfertigt werden können: Gemäß § 7 der „Verordnung über die Nutzenbewertung von Arzneimitteln nach § 35a Absatz 1 SGB V für Erstattungsvereinbarungen nach § 130b SGB V" soll bei der Nutzenbewertung geprüft werden, „ob für das Arzneimittel ein Zusatznutzen gegenüber der zweckmäßigen Vergleichstherapie belegt ist, […] und mit welcher Wahrscheinlichkeit der Beleg jeweils erbracht wird." Diese Aufgabe obliegt dem IQWiG, das dazu Methoden der evidenzbasierten Medizin anwendet, die in den Allgemeinen Methoden [2] beschrieben werden. Aus verschiedenen Gründen ist es in der Vergangenheit vorgekommen, dass ein Zusatznutzen nur für eine Subgruppe einer Studie festgestellt wurde.

Die weitere Einschränkung erscheint dann, wenn sie anhand der gleichen Studiendaten begründet wird, auffällig. Viele Unterschiede in den Entscheidungsstrategien bestehen dabei nur scheinbar, weil die meisten regulatorischen Guidelines den formalen Nachweis der Wirksamkeit thematisieren, während die Methoden des IQWiG ein Regelwerk für die Bewertung des Nutzens und speziell des Zusatznutzens darstellen. Auch im Rahmen der Zulassung wird nach der Bewertung der Wirksamkeit zur Feststellung eines prinzipiell positiven Nutzen-Risiko-Verhältnisses der Unterschied zwischen den Therapiegruppen anhand verschiedener Variablen betrachtet, die dann im Rahmen der Feststellung eines Zusatznutzens weiter ausspezifiziert und bewertet werden.

Weitere Unterschiede ergeben sich dann, wenn die im Rahmen der Zulassungsstudien gewählte Vergleichstherapie nicht als zweckmäßige Vergleichstherapie eingestuft wird und folglich im Rahmen der Nutzenbewertung vergleichende Evidenz nur für eine Subpopulation der Studienpopulation akzeptiert wird.

In der folgenden Diskussion möchten wir uns jedoch auf die Situation beschränken, in der die Entscheidung für eine Einschränkung der Population mit den Ergebnissen der Studie in der Gesamtpopulation und in einer nach bestimmten Kriterien ausgewählten Untergruppe begründet wird. Wir betrachten Situationen, in denen sowohl die für die Zulassung ge-

wählte Vergleichstherapie als auch der für die Zulassung gewählte Endpunkt für die Nutzenbewertung entscheidungsrelevant sind.

Kardiologie und Onkologie liefern die Beispiele für Indikationen, in denen häufig die Bewertung der Wirksamkeit und des Nutzens an den gleichen Endpunkten gemessen wird. Damit schränken wir die Diskussion auf Situationen ein, in denen letztlich trotz eines „signifikanten" Gesamtergebnisses eine gewisse Unzufriedenheit mit der Relevanz des Studienergebnisses im Sinne einer detaillierten Nutzen-Risiko-Bewertung ausschlaggebend für eine Einschränkung auf eine nicht präspezifizierte Untergruppe der Studienpopulation ist.

Unsere Diskussion beschränkt sich auch auf die Diskussion einer Entscheidungssituation, die im Wesentlichen dem Stand am Ende der Arzneimittelentwicklung entspricht, wo gerade in den vorgenannten Indikationen häufig nur eine Phase-III-Studie vorliegt.

Lehrbuchbeispiele der Diskussion um die Bedeutung der Ergebnisse aus Untergruppen einer Studie liefern CAPRIE, IPASS und PLATO:

Die CAPRIE (Clopidogrel versus Aspirin in Patients at Risk of Ischemic Events) Studie war eine randomisierte, doppelblinde, multizentrische Studie zum Vergleich von Clopidogrel versus Aspirin bei Patienten mit vorausgegangenen Verschlusserkrankungen. Die Randomisierung erfolgte stratifiziert für die Untergruppen vorausgegangener Schlaganfall, vorausgegangener Myokardinfarkt (MI) oder bestehende periphere arterielle Verschlusskrankheit (pAVK). Der zusammengesetzte primäre Endpunkt war der Anteil der Patienten, die einen ischämischen Schlaganfall, Myokardinfarkt oder vaskulären Tod erlitten haben. Die relative Risikoreduktion war in der Gesamtpopulation zugunsten von Clopidogrel signifikant erhöht, jedoch war der Therapieeffekt in den Untergruppen nicht konsistent. Während in der pAVK-Untergruppe eine signifikante Risikoreduktion zugunsten von Clopidogrel zu beobachten war, zeigte sich in der MI-Gruppe ein numerisch gegenteiliger Therapieeffekt (zugunsten von Aspirin) [3]. Dies hat im Rahmen des Zulassungsprozesses zu substanziellen Diskussionen geführt, bevor eine Zulassung für die Gesamtpopulation ausgesprochen wurde, da zumindest die Nichtunterlegenheit von Clopidogrel gegenüber ASS als gegeben angesehen wurde. Im Anschluss an diese Diskussion hat das IQWiG einen Zusatznutzen lediglich für die Patienten mit einer peripheren Verschlusskrankheit anerkannt [4, 5].

Eine Einschränkung der Zulassung auf eine relevante Untergruppe wurde in der IPASS-Studie (Iressa Pan Asia Study) vorgenommen. In dieser randomisierten Phase-III-Studie [6] sollte die Überlegenheit von Gefinitib gegenüber der Standardtherapie (Carboplatin und Paclitaxel) bei Nichtrauchern und Wenigrauchern mit kleinzelligem Bronchialkarzinom gezeigt werden. Der primäre Endpunkt war das progressionsfreie Überleben. Obwohl der Therapieeffekt signifikant war, erschwerten sich kreuzende Überlebenskurven die Interpretation der Studie. Daraufhin wurde bei Patienten mit einer EGFR (epidermal growth factor) -Mutation entdeckt, dass unter Gefitinib das progressionsfreie Überleben signifikant länger war. In der Untergruppe ohne eine EGFR-Mutation war hingegen die bisherige Standardtherapie überlegen, sodass (nach einer Bestätigungsstudie) die Zulassung auf die Untergruppe mit EGFR-Mutation beschränkt wurde.

Der Trend hin zu klinischen Studien, die nicht nur multizentrisch, sondern in verschiedenen Regionen der Welt durchgeführt werden, hat die Frage der Beurteilung von Konsistenz der Studienergebnisse in den einzelnen Regionen in den Mittelpunkt der Diskussion gerückt. PLATO (Platelet Inhibition and Patient Outcomes), eine multiregionale, multizentrische, randomisierte Phase-III-Doppelblindstudie, hat eine deutliche Überlegenheit von Ticagrelor gegenüber Clopidogrel bei Patienten mit akutem Koronarsyndrom belegt. Die sorgfältige Prüfung des Studienergebnisses hat jedoch auch ergeben, dass der Behandlungseffekt in Nordamerika in entgegengesetzter Richtung einen Vorteil für die Behandlung mit Clopidogrel aufwies. Auch wenn Zufall als mögliche Ursache nie ausgeschlossen werden kann, haben sich die Patienten in den verschiedenen Regionen dadurch unterschieden, dass in Nordamerika traditionell deutlich mehr Patienten eine hohe Dosis Aspirin als Begleitmedikation erhalten hatten, die dann in der Folge in Kombination mit Ticagrelor kontraindiziert wurde [7].

Die Biometrie ist lange Zeit in der Medizin dafür eingetreten, dass in einer klinischen Studie eine klare Hypothese mit einem primären Endpunkt in einer präspezifizierten (Intention-to-Treat) Population untersucht werden sollte und nur dies eine Kontrolle des Fehlers einer falsch-positiven Entscheidung (Fehler 1. Art) bezüglich der Wirksamkeit einer Therapie erlaubt. Auch wenn die Konzepte des multiplen Testens eine größere Freiheit in Bezug auf die primäre Auswertungsstrategie ermöglichen, ist die genaue Planung die Voraussetzung für die Gewinnung glaubwürdiger Schlussfolgerungen. „Fishing for significance" und „data dredging" beschreiben die Zurückhaltung gegenüber Ergebnissen, die aus nicht genau präspezifizierten Auswertungen folgen. Es ist wohlbekannt, dass man mit hoher Wahrscheinlichkeit wenigstens einen „signifikanten" Therapieunterschied finden wird, wenn man zum Beispiel den Therapieeffekt in 20 Untergruppen der Patientenpopulation einer Studie untersucht, auch wenn die Therapie in Wirklichkeit genauso wirksam wie Placebo ist. Gemäß den europäischen Richtlinien der Arzneimittelentwicklung ist eine Replikation des Ergebnisses dringend erforderlich, wenn die Untersuchungen der Untergruppe nicht als Teil einer multiplen Teststrategie vorgeplant gewesen sind [8].

Ebenso klar ist es, dass man auch bei einer eigentlich wirksamen Therapie bei der Untersuchung von Untergruppen rein zufällig einmal keinen Unterschied zwischen den Therapiegruppen finden wird. Daher sollte dieselbe Zurückhaltung geübt werden, wenn man wie in den oben beschriebenen Untersuchungen von Untergruppen der Patientenpopulation ein Ergebnis findet, das nicht der global begründeten Erwartung an den Therapieeffekt entspricht. Eine Replikation dieses Untergruppenbefunds (oder der Nachweis, dass es sich nicht lediglich um einen zufällig beobachteten reduzierten Therapieeffekt handelt) kann unter Umständen ethisch schwer zu vertreten sein, wenn das Nebenwirkungsprofil des Arz-

Bundesgesundheitsbl 2015 · 58:274–282 DOI 10.1007/s00103-014-2105-2

A. Gonnermann · M. Kottas · A. Koch

Biometrische Entscheidungsunterstützung in Zulassung und Nutzenbewertung am Beispiel der Implikationen von heterogenen Ergebnissen in Untergruppen der Studienpopulation

Zusammenfassung

Sowohl im Rahmen der Zulassung neuer Arzneimittel als auch im Rahmen der Bewertung eines Zusatznutzens zur Rechtfertigung höherer Kosten findet gelegentlich eine Einschränkung der Zielpopulation auf eine Untergruppe der Studienpopulation statt. Bei der Zulassung geht es dann häufig darum, dass in einer Untergruppe das Nutzen-Risiko-Verhältnis als nicht positiv eingeschätzt wird. In der Diskussion um die Erstattungsfähigkeit geht es um die Feststellung eines Zusatznutzens, der möglicherweise nicht für die gesamte Studienpopulation anerkannt werden kann. In seinen Bewertungskriterien verweist das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) auf publizierte Arbeiten, die Kriterien für die Bewertung der Glaubwürdigkeit von Wirksamkeitsbehauptungen auf der Basis von Ergebnissen in Untergruppen einer Studienpopulation benennen (BMJ 340:850–854, 2010). Viele der dort genannten Kriterien haben Eingang in die regulatorische Diskussion gefunden, die nun im Entwurf einer kürzlich veröffentlichten Guideline der European Medicines Agency (EMA) niedergelegt ist. Als ein Kriterium für die Glaubwürdigkeit eines Ergebnisses in einer Untergruppe der Studienpopulation wird die Signifikanz des Tests auf Interaktion oder Heterogenität genannt. In dieser Arbeit wird die Sinnhaftigkeit dieser Forderung kritisch hinterfragt. Wir vertreten die Auffassung, dass dieses Kriterium nicht geeignet ist, eine formale Entscheidung darüber herbeizuführen, ob ein Therapieeffekt in einer Untergruppe glaubwürdig ist oder nicht. Wir führen hierzu methodische und erkenntnislogische Argumente an, die, nach unserer Auffassung, sowohl bei der Bewertung der Wirksamkeit eines Arzneimittels als auch bei der Nutzenbewertung Gültigkeit haben. Wir sehen dies als ein Beispiel dafür, dass dieselben Entscheidungsstrategien in Zulassung und Nutzenbewertung eingesetzt werden sollten, auch wenn die Bewertung der Daten im Einzelfall in Abhängigkeit von durchaus unterschiedlichen Zielen zu unterschiedlichen Schlussfolgerungen führen kann.

Schlüsselwörter

Subgruppen · Heterogenität · Inkonsistenz · Metaanalyse · Nutzen-Risiko-Bewertung

Biostatistical support for decision making in drug licensing and reimbursement exemplified by implications of heterogeneous findings in subgroups of the study population

Abstract

In the context of both drug licensing and reimbursement, the target population is sometimes restricted to a specific subgroup. In the setting of drug licensing the discussion concerns a negative benefit/risk assessment in a relevant subgroup. For reimbursement the debate involves the detection of an additional benefit compared with standard treatment, which can in some situations not be accepted for the overall study population. In their Methods Paper, the Institute for Quality and Efficiency in Health Care (IQWiG) refers to published articles that name criteria for the evaluation of credibility to claim a therapeutic effect on the basis of results in the subgroups of a study population (BMJ 340:850–854, 2010). A number of these criteria have found their way into the regulatory debate, which was recently published in a draft guideline of the European Medicines Agency (EMA). However, the significance of the interaction/heterogeneity test has been mentioned as one criterion for the credibility of a finding in a subgroup of the study population. This aspect is critically challenged in our paper. In our estimation, the application of this criterion hinders the critical discussion of whether a global treatment effect is applicable to relevant subgroups of a study population and the potential implications of this. We feel that biostatistical support for decision-making strategies should be the same in both worlds, even though in some instances the outcomes in a specific situation may be different, depending on the objective to be demonstrated.

Keywords

Subgroups · Heterogeneity · Inconsistency · Meta-analysis · Benefit/risk-assessment

neimittels bei einem geringen Therapieeffekt ein positives Nutzen-Risiko-Verhältnis infrage stellt. Aus diesem Grund bleibt im Rahmen der Arzneimittelzulassung oft als einziger Ausweg, die Anwendung des Arzneimittels in einer Untergruppe nicht zu empfehlen (Einschränkung der Indikation).

Die gleiche Zurückhaltung wäre dann auch erforderlich, wenn man die Bewertung eines Zusatznutzens in einer nicht präspezifizierten Untergruppe zur Basis einer Entscheidung im Kontext Nutzenbewertung machen möchte. Die Glaubwürdigkeit des Subgruppenbefunds muss geeignet eingeschätzt werden.

In einer früheren Arbeit haben wir argumentiert, dass bei Anwendung, aus Sicht der Biometrie, gleicher Entscheidungsregeln in der Bewertung der Wirksamkeit und des Zusatznutzens aufgrund unterschiedlicher Ziele unterschiedliche Schlussfolgerungen gezogen werden (Bewertung der Wirksamkeit auf der Basis von Nichtunterlegenheit, Bewertung eines Zusatznutzens, der durch bessere Wirksamkeit begründet werden soll als prominentestes Beispiel einer derartigen Entscheidungssituation) [4].

Wohl wissend um die Unterschiede zwischen Wirksamkeitsnachweis und Nutzenbewertung möchten wir in dieser Arbeit aus Sicht der Zulassung von Arzneimitteln dafür argumentieren, dass, da es letztlich um die Bewertung von Evidenz auf der Basis von Daten geht, in beiden Entscheidungssituationen die gleichen Entscheidungsstrategien zur Anwendung kommen sollten. Methodische Überlegungen um die Glaubwürdigkeit

von Ergebnissen in Untergruppen einer Studie sollen dabei als Beispiel dienen.

Wenig ist bisher dazu gesagt worden, ab wann ein Behandlungseffekt in einer Untergruppe als „zu klein" eingestuft oder ganz allgemein die Heterogenität der Behandlungseffekte in Untergruppen der Studienpopulation als so groß eingestuft werden sollte, dass eine separate Nutzen-Risiko-Bewertung sowie die Diskussion der Implikationen erforderlich ist. In dieser Arbeit untersuchen wir verschiedene, teilweise empirisch begründete Methoden zur Generierung von Signalen dafür, dass ein Globaleffekt möglicherweise nicht auf die gesamte Patientenpopulation in gleicher Weise angewendet werden kann. Dazu werden in einer Simulationsstudie die empirischen Raten an falsch-positiven und falsch-negativen Entscheidungen verglichen, wenn es um die Frage geht, ob Heterogenität zwischen Untergruppen vorliegt.

In einer Serie von Arbeiten haben Sun et al. [1, 9–11] als Fortführung der Arbeit von Oxmann und Guyatt [12] Kriterien aufgestellt, wann die Ergebnisse von Subgruppenanalysen als glaubwürdig eingestuft werden sollten. Angesprochen wird, dass in einer derartigen Bewertung praktisch nie alle Kriterien vollständig erfüllt sein können, sondern dass, wie auch bei der Indikationsstellung für eine Therapie, verschiedene Einzelbewertungen in ein Gesamturteil bezüglich der Eignung/Glaubwürdigkeit einmünden. Auf diese Arbeiten beziehen sich auch die Allgemeinen Methoden des IQWiG [2] und viele der genannten Kriterien haben Eingang in den Entwurf der europäischen Guideline zur Bewertung von Untergruppen in klinischen Studien der Phase III (Guideline on the investigation of subgroups in confirmatory clinical trials [13]) gefunden.

Sun et al. [1] und vorausgegangene Arbeiten haben die Signifikanz des Tests auf Interaktion als eines der Kriterien genannt, von denen die Glaubwürdigkeit eines Behandlungseffekts in Untergruppen abhängig gemacht werden soll. Kriterium 7 in der entsprechenden Liste lautet: „Does the interaction test suggest a low likelihood that chance explains the apparent subgroup effect?" (Seite 851).

Vor dem Hintergrund der Ergebnisse unserer Simulationsstudie diskutieren wir kritisch, ob dieses Kriterium sinnhaft im Lichte der Realität von Entscheidungsnotwendigkeiten im Bereich der Zulassung und der Bewertung des Zusatznutzens eingesetzt werden kann, in der am Ende eine Ja/Nein-Entscheidung über die Bewertung eines Ergebnisses in einer Untergruppe zu fällen ist. Nicht nur weil dies in aktuellen Diskussionen so vorgetragen wurde, sondern auch um die Bedeutung dieses Kriteriums genau herauszuarbeiten, betrachten wir dabei den Fall, dass der Heterogenitätstest im Sinne eines Gatekeepers für die Glaubwürdigkeit eines Untergruppeneffekts genommen wird, auch wenn uns bewusst ist, dass medizinische Überlegungen zur Plausibilität der gefundenen Untergruppeneffekte und viele der anderen in Sun et al. [1] genannten Kriterien zusätzlich in eine Gesamtbewertung mit einbezogen werden müssen.

Methoden

Für die Beurteilung der Konsistenz der Therapieeffekte in den Untergruppen und somit die Anwendbarkeit des (globalen) Therapieeffekts auf alle relevanten Untergruppen gibt es verschiedene Interaktions- bzw. Heterogenitätstests. Chen et al. [14] haben eine Übersicht über Methoden publiziert, die verwendet werden können, um die Konsistenz der Therapieeffekte in multiregionalen Studien zu beurteilen. Die Autoren untersuchen auch eine Regel aus der japanischen MHLW (Ministry of Health, Labour and Welfare) Richtlinie aus dem Jahre 2007. Dort wird Konsistenz so definiert, dass der in einer Region beobachtete Therapieeffekt (z. B. bei Patienten, die in Japan eingeschlossen wurden) im Verhältnis zum Therapieeffekt in der gesamten Studienpopulation mindestens halb so groß (50 %) wie der Behandlungseffekt in der gesamten Studienpopulation sein soll. Quan et al. [15] haben diese Regel auf mehrere Regionen erweitert.

Weitere Vorschläge finden sich in Gail und Simon [16] und Piantadosi und Gail [17], die einen standardisierten Range Test vorschlagen, den sie mit dem Gail-und-Simon-Test vergleichen. Der Range Test hat im Vergleich zum Gail-und-Simon-Test eine größere Power, wenn die Behandlungseffekte nur in wenigen Untergruppen ein unterschiedliches Vorzeichen haben (d. h. die neue Behandlung in einer Untergruppe dem bisherigen Standard unterlegen ist). Liegen jedoch mehrere Untergruppeneffekte in entgegengesetzter Richtung, hat der Gail-und-Simon-Test eine größere Power. Li und Chan [18] schlagen einen erweiterten Range Test vor, wobei alle beobachteten Behandlungsdifferenzen betrachtet werden und nicht nur das Minimum und Maximum wie beim standardisierten Range Test. Ihre Simulationen zeigen, dass der erweiterte Range Test generell eine größere Power hat als der Range Test von Piantadosi und Gail. Im Vergleich zum Gail-und-Simon-Test hat der erweiterte Range Test auch eine größere Power, wenn eine Behandlung in den meisten Subgruppen überlegen ist, andernfalls verliert der Test nicht viel Power.

Der Breslow-Day-Test auf Heterogenität ist der sicherlich am weitesten verbreitete Test zur Untersuchung der Konsistenz eines Therapieeffekts in verschiedenen Untergruppen einer Studienpopulation [19]. In einer Simulationsstudie vergleichen Bagheri et al. [19] den Breslow-Day-Test mit einem Likelihood-Quotienten-Test basierend auf einem gemischten logistischen Modell sowie mit Cochran's Q bei multizentrischen Studien. Sie zeigen, dass eine ungleiche Aufteilung der gesamten Studienpopulation über die Zentren die Power der Tests im Vergleich zu der Situation, in der in allen Zentren die gleiche Patientenzahl rekrutiert wird, reduziert. Der Fehler 1. Art ist in allen Tests nahe dem nominalen Signifikanzniveau von 5 %. In dieser Studie hat der Breslow-Day-Test die größte Power.

All diesen Methoden ist gemein, dass sie eine geringe Power haben. Zu bedenken ist auch, dass Heterogenitätstests quasi „verkehrt herum" testen: Die Nullhypothese dieser Tests ist, dass es einen konsistenten Therapieeffekt in den Untergruppen gibt. Klassischerweise ist das Ziel beim statistischen Testen die Nullhypothese abzulehnen. Bei der Konsistenzbewertung in klinischen Studien soll jedoch gerade diese Nullhypothese bestätigt werden, um so zu rechtfertigen, dass eine Therapie in allen relevanten Untergruppen angewendet werden kann.

Unsere Untersuchungen beschränken sich auf zweiarmige klinische Studien mit

einem dichotomen Endpunkt. Zur Generierung von Signalen für die Inkonsistenz von Studienergebnissen werden vier Strategien untersucht und mit Cochran's Q, der dem Breslow-Day-Test entspricht, verglichen. Dabei wird die Konsistenz des Behandlungseffekts in Untergruppen der Patientenpopulation infrage gestellt, wenn (für eine Studie in der ein signifikanter Therapieeffekt beobachtet wurde) jeweils eine der folgenden Regeln erfüllt ist:

1. Q-Regel: Der *p*-Wert für Cochran's Q ist kleiner als 0,15.
2. G-(Guideline) Regel: Der Behandlungseffekt in einer Subgruppe ist entweder halb oder doppelt so groß wie der Gesamtbehandlungseffekt. Dieser Ansatz wurde als empirische Regel in einer Arbeitsversion des zwischenzeitlich publizierten Entwurfs der Subgruppen-Guideline der EMA zur Bewertung von Untergruppenergebnissen in konfirmatorischen klinischen Studien der Phase III erwähnt [13] und ist strukturell ähnlich zur Regel von Quan et al. [15].
3. E-(Epidemiologie) Regel: Der Behandlungseffekt in einer Subgruppe ist entweder ein Viertel oder viermal so groß wie der Gesamtbehandlungseffekt.
4. KI-(Konfidenzintervall) Regel: Der Punktschätzer einer Subgruppe liegt nicht im Konfidenzintervall des Gesamtbehandlungseffektes. Diese empirische Regel wird im aktuellen Entwurf der Guideline zur Untersuchung von Subgruppen erwähnt [13].

Um die technischen Eigenschaften dieser Regeln zu untersuchen, sind wir von einer Studie ausgegangen, die durch eine Kovariable (z. B. Geschlecht) in zwei Subgruppen geteilt wird. Für den Gesamtbehandlungseffekt der Studie wird ein Odds Ratio von 1,5 festgelegt, sodass unter der Annahme einer Erfolgsrate von 0,2 im Kontrollarm und einer Erfolgsrate von 0,27 im Therapiearm unter Berücksichtigung eines Fehlers 1. Art von 5 % und einer Power von 80 % insgesamt 1150 Patienten rekrutiert werden müssen. Während die Aufteilung in Kontroll- und Interventionsgruppe in einer 1:1 Randomisierung erfolgte, wurde das Verhältnis der Untergruppen zueinander von 50 zu 50 % schrittweise bis zu einer Aufteilung 90 zu 10 % verändert.

Die grundsätzliche Annahme in unserer Simulationsstudie ist, dass in den beiden Untergruppen jeweils ein Modell mit festen Effekten (Fixed Effect Modell, FEM) vorliegt. Für Simulationen im FEM wurden Wahrscheinlichkeiten zur Erzeugung der binomial verteilten Anzahl an Erfolgen direkt vorgegeben: $X_{ij} \sim Bin(n_{ij}; p_{ij})$. Wird ein konsistenter Therapieeffekt in den beiden Untergruppen angenommen, so befinden wir uns unter der Nullhypothese (H_0) des Tests auf Heterogenität. Zur Simulation wurden für die beiden Untergruppen die gleichen Erfolgswahrscheinlichkeiten und Therapieeffekte angenommen. Dabei entspricht der Anteil der Simulationsdurchläufe, bei denen fälschlicherweise ein Signal für inkonsistente Ergebnisse in Untergruppen erfolgte, dem Anteil der falsch-positiven Testergebnisse (Fehler 1. Art bzw. im Sinne eines diagnostischen Tests: 1-Spezifität). Heterogenität entsteht in diesem Modell, indem verschiedene Erfolgswahrscheinlichkeiten und dadurch verschiedene Therapieeffekte in den beiden Untergruppen angenommen werden.

Es wurden zwei Alternativhypothesen für den Test auf Heterogenität untersucht: Als Erstes wurden die Wahrscheinlichkeiten für die Simulation der Anzahl an Erfolgen so eingestellt, dass in einer Untergruppe der Therapieeffekt doppelt (oder halb) so groß ist wie der Therapieeffekt in der Gesamtpopulation. Somit trifft genau die Situation zu, die als „Guideline Regel" beschrieben ist und die Homogenität der Ergebnisse infrage stellen sollte (H_{1G}). Als zweite Alternativhypothese im FEM ist der Effekt in einer Untergruppe viermal (oder ein viertel Mal) so groß im Vergleich zum Gesamtbehandlungseffekt: Die sogenannte Epidemiologie-Regel trifft zu (H_{1E}).

Im Bereich von Metaanalysen wird für die Entstehung von Heterogenität zwischen Studien das Modell mit zufälligen Effekten (Random Effects Modell, REM) nach DerSimonian & Laird diskutiert [20]. Wir haben zur Bewertung der Regeln daher auch in diesem Modell eine Simulationsstudie durchgeführt. Für das REM nach DerSimonian & Laird werden in einem zweistufigen Zufallsprozess zunächst normal verteilte Logits gezogen ($\text{logit}(p_{ij}) = p_{ij} / (1 - p_{ij}) \sim N(\mu_{ij}; \sigma^2_{ij})$, wobei i für die Anzahl der Subgruppen und j für die Interventions- bzw. und Kontrollgruppe steht). Aus den Logits werden durch Rücktransformation die Erfolgswahrscheinlichkeiten in der Therapie- und Kontrollgruppe berechnet $p_{ij} = 1 / (1 + \exp(-\text{logit}(p_{ij})))$. Diese Wahrscheinlichkeiten werden verwendet, um binomialverteilte Anzahlen der Erfolge zu einer vorgegebenen Patientenanzahl zu simulieren, die die Ergebnisse in den Untergruppen darstellen. Diese Simulationen verbinden die hier angestellten Überlegungen zur Generierung von Signalen für Heterogenität zu einer früheren Arbeit, in der wir die Auswirkungen von Heterogenität in einem REM auf die Validität von Schlussfolgerungen in Studien diskutiert haben, die mehrere Regionen/Ethnizitäten einschließen [21].

Unter der Nullhypothese des Heterogenitätstests haben alle Subgruppen die gleichen (d. h. konsistenten) Therapieeffekte. Für die Simulation im REM wird dafür σ^2 auf 0 gesetzt. Unter der Alternativhypothese sind die Therapieeffekte in den Subgruppen nicht mehr konsistent. Dazu wurde σ^2 so eingestellt, dass das Heterogenitätsmaß I^2 ca. 25 % ($H_{1;25\%}$) bzw. ca. 50 % ($H_{1;50\%}$) beträgt. Sowohl im FEM als auch im REM entspricht die Rate an abgelehnten Nullhypothesen unter der Alternativhypothese (es bestehen inkonsistente Therapieeffekte in den Untergruppen) von Cochran's Q und den übrigen Methoden der Power bzw. im Sinne eines diagnostischen Tests der Sensitivität der Regel.

Alle Simulationen sind mit 10.000 Replikationen durchgeführt worden.

Ergebnisse

Tab. 1 zeigt den empirischen Fehler 1. Art (1-Spezifität) und die empirische Power (Sensitivität) der untersuchten Regeln für zwei Subgruppen in dem Modell mit festen Effekten (FEM). Die Untergruppen wurden in drei verschiedene Verhältnisse aufgeteilt, wobei der Anteil der Personen in Interventions- und Kontrollgruppe im Sinne einer 1:1 Randomisierung gleich geblieben ist. Während bei I^2 der Mittelwert der einzelnen Simulationsergebnisse und

Tab. 1 Empirischer Fehler 1. Art bzw. Power bei Annahme von festen Effekten in den einzelnen Strata bei zwei Untergruppen

Situation	Verhältnis	I^{2a}	Cochran's Q	Q-Regel	G-Regel	E-Regel	KI-Regel
H_0	50:50	0,1489	0,1449	0,1427	0,0723	0,0179	0,0446
H_{1G}	50:50	0,4192	0,4804	0,4798	0,4580	0,2232	0,2998
H_{1E}	50:50	0,6136	0,7192	0,7172	0,7006	0,4395	0,5656
H_0	70:30	0,1495	0,1494	0,1541	0,2014	0,0725	0,2071
H_{1G}	70:30	0,3093	0,3472	0,3480	0,5110	0,3321	0,4299
H_{1E}	70:30	0,4374	0,5018	0,5027	0,6753	0,4927	0,6003
H_0	90:10	0,1486	0,1474	0,1459	0,4580	0,2237	0,5127
H_{1G}	90:10	0,2183	0,2356	0,2370	0,6176	0,4545	0,6100
H_{1E}	90:10	0,2668	0,2909	0,2943	0,6831	0,5496	0,6742

[a]Bei I^2 handelt es sich um den Mittelwert über alle Simulationen

Tab. 2 Empirischer Fehler 1. Art bzw. Power bei Annahme von zufälligen Effekten in den einzelnen Strata bei zwei Untergruppen

Situation	Verhältnis	I^{2a}	Cochran's Q	Q-Regel	G-Regel	E-Regel	KI-Regel
H_0	50:50	0,1489	0,1449	0,1427	0,0723	0,0179	0,0446
$H_{1;25\,\%}$	50:50	0,2559	0,2784	0,2744	0,1333	0,0433	0,1340
$H_{1;50\,\%}$	50:50	0,5093	0,5684	0,5605	0,2933	0,1329	0,4254
H_0	70:30	0,1495	0,1494	0,1541	0,2014	0,0725	0,2071
$H_{1;25\,\%}$	70:30	0,2355	0,2519	0,2477	0,2652	0,1156	0,3101
$H_{1;50\,\%}$	70:30	0,4702	0,5312	0,5252	0,4488	0,2365	0,5701
H_0	90:10	0,1486	0,1474	0,1459	0,4580	0,2237	0,5127
$H_{1;25\,\%}$	90:10	0,1868	0,1929	0,1862	0,4702	0,2430	0,5470
$H_{1;50\,\%}$	90:10	0,3424	0,3836	0,3854	0,5729	0,3538	0,6911

[a]Bei I^2 handelt es sich um den Mittelwert über alle Simulationen

bei Cochran's Q (welcher dem Breslow-Day-Test entspricht) die durchschnittliche Rate abgelehnter Nullhypothesen auf Basis aller Simulationsdurchläufe berechnet wurden, bilden die Regeln die durchschnittliche Rate abgelehnter Nullhypothesen auf Basis aller signifikanten Therapieeffekte ab. Dadurch lässt sich der empirische Fehler 1. Art (bzw. die Power) von Cochran's Q mit dem empirischen Fehler 1. Art (bzw. der Power) der Q-Regel bei Restriktion auf signifikante Therapieeffekte vergleichen.

Bei einem Untergruppenverhältnis von 50:50 erreicht die Q-Regel maximal 71,72 % und hat im Vergleich zu den anderen Regeln die höchste Power. Zwar haben die anderen Regeln bei einer 50:50 Aufteilung einen geringeren Fehler 1.Art als die Q-Regel, aber dafür ist auch ihre Power geringer. Die G-Regel hat eine ähnliche Power (70,06 %) und der Fehler 1. Art ist kleiner (7,23 %), was einen klaren Vorteil dieses Ansatzes darstellen würde. Die Q-Regel hält jedoch in allen Szenarien, d. h. auch bei ungleichmäßiger Aufteilung der Untergruppen, den Fehler 1. Art bei ca. 15 % ein. Bei den anderen Regeln steigt der Fehler 1. Art an, je extremer die Aufteilung der Untergruppen ist.

In **Tab. 2** sind die Ergebnisse für zwei Untergruppen im REM dargestellt.

In einem Modell mit zufälligen Effekten mit zwei Untergruppen haben in 50:50 aufgeteilten Untergruppen alle Regeln eine geringe Power (mit Ausnahme von Cochran's Q und der Q-Regel). Wie zuvor hat die Q-Regel in diesem Fall die größte Power mit maximal 56,05 %. Bei G-, E- und KI-Regel ist der Fehler 1. Art geringer als bei der Q-Regel, aber auch die Power ist niedriger. Bei ungleichmäßiger Aufteilung der Untergruppen verliert die Q-Regel an Power, während der Fehler 1. Art der anderen Regeln steigt. Qualitativ unterscheiden sich die Ergebnisse von REM und FEM nicht voneinander.

Diskussion

Randomisierte doppelblinde Studien zum Nachweis der Wirksamkeit und eines positiven Nutzen-Risiko-Verhältnisses sind der Goldstandard für die Zulassung neuer Arzneimittel. In Phase III sollen diese Studien in aller Regel eine weit gefasste Patientenpopulation einschließen, sodass im Erfolgsfall eine möglichst wenig eingeschränkte Patientenpopulation behandelt werden kann. Es ist jedoch eine nicht triviale Annahme, dass alle Patienten in gleichem Maße von der zu prüfenden Therapie profitieren und folglich ist die Annahme eines konsistenten Therapieeffekts in relevanten Untergruppen der Patientenpopulation zu prüfen.

Auch im Rahmen der Bewertung des Zusatznutzens von Arzneimitteln, der höhere Kosten für das Arzneimittel rechtfertigen würde, spielt die Konsistenz der Behandlungseffekte in relevanten Untergruppen neben anderen Kriterien (wie z. B. der zweckmäßigen Vergleichstherapie, die für bestimmte Patientenpopulationen unterschiedlich definiert sein mag) eine wesentliche Rolle. Seitens der verantwortlichen Stellen besteht darüber hinaus eine gewisse Bereitschaft, einen Zusatznutzen eventuell auch nur für eine Untergruppe der initialen Studienpopulation zu bestätigen.

Bisher ist der Begriff Konsistenz in der Literatur nicht klar definiert und operationalisiert. Forest-Plots und Homogeni-

tätstests werden benutzt, um die Homogenität des Behandlungseffekts in relevanten Untergruppen plausibel zu machen oder die Konsistenz der Ergebnisse infrage zu stellen. In dieser Arbeit haben wir eine Reihe von Regeln untersucht, mit deren Hilfe man ein Signal für das Vorliegen inkonsistenter Therapieeffekte in Untergruppen der Patientenpopulationen operationalisieren kann und mit Cochran's Q, quasi als dem Standard zur Untersuchung dieser Frage, verglichen. Andere Möglichkeiten zur Feststellung der Heterogenität wurden in der Literatur bzw. in dem Entwurf zur neuen Guideline der EMA erwähnt. Wir haben Simulationen unter der Annahme eines Modells mit festem Effekt und unter der Annahme eines Modells mit zufälligen Effekten durchgeführt. Es wurde die Situation betrachtet, dass die Studienpopulation in zwei gleich große oder zwei unterschiedlich große Untergruppen zerfällt. Die Fallzahl wurde so gewählt, dass der Therapieeffekt in der Gesamtpopulation mit einer Wahrscheinlichkeit von 80 % entdeckt werden kann. Damit werden genau die Verhältnisse abgebildet, die in klinischen Studien zum Wirksamkeitsnachweis bezüglich der Entdeckung und Bewertung von Heterogenität vorliegen, die üblicherweise bei der Studienplanung unberücksichtigt bleiben.

In der Summe ist zu konstatieren, dass über den Gesamtbereich möglicher Aufteilungen zwischen Patienten mit und ohne Risikofaktor kein uniform optimales Vorgehen ausgezeichnet werden kann. Es wird vermutlich nur möglich sein, für bestimmte Situationen mehr oder weniger geeignete Regeln zu identifizieren, wann die Konsistenz der Therapieeffekte in den Untergruppen infrage gestellt werden sollte.

Die Q-Regel, die der Anwendung des Breslow-Day-Tests für die Therapievergleiche in den Untergruppen einer Studie entspricht, in der der Therapieeffekt signifikant gewesen ist, hat dabei den Vorteil, dass ein vorher festgelegtes Fehlerniveau eingehalten wird. Die vorgestellten Ergebnisse bestätigen erneut die geringe Power dieses Tests sogar unter der Bedingung, dass der Fehler 1. Art für diesen Test auf 15 % erhöht wird. Der Powerverlust ist umso dramatischer, je ungleicher sich die Patienten auf die Untergruppen verteilen. Dies ist eine Folge davon, dass gewichtete Abweichungen der Schätzungen für den Therapieeffekt in den Untergruppen vom Gesamtschätzer die Grundlage der Berechnung der Teststatistik bilden. Selbst wenig spektakuläre Abweichungen (wie beispielsweise das Geschlechterverhältnis in Herzinfarktstudien, das üblicherweise in der Größenordnung 70 %/30 % „zugunsten" männlicher Studienteilnehmer verschoben ist) führen dazu, dass die Power des Breslow-Day-Tests teilweise kleiner als 50 % ist. Dies gilt sogar dann, wenn ein vierfacher Unterschied des Therapieeffekts in den Untergruppen im Vergleich zum Gesamteffekt entdeckt werden soll. Die G- bzw. die E-Regel sind in Abhängigkeit der Anzahl der Untergruppen bei ungleichmäßigen Aufteilungen besser geeignet, wenn akzeptiert wird, dass die Rate falsch-positiver Schlussfolgerungen nicht begrenzt ist.

Ein Signal darüber zu definieren, dass in einer kleinen Untergruppe ein extremer (z. B. vierfach erhöhter) Effekt beobachtet wird, ist gängige Praxis im Bereich der Arzneimittelsicherheitsbewertung. Dies wird dadurch begründet, dass in einer kleinen wohldefinierten Gruppe auch eine extreme Risikoerhöhung nicht signifikant werden kann. Die höhere Trennschärfe dieser Regel ist mit dem Preis verbunden, dass eine höhere Anzahl falsch-positiver Befunde generiert wird.

Die Diskussion um den Stellenwert der Heterogenitätstests und Konsistenzuntersuchungen ist keineswegs neu, sondern beschäftigt die Forschung im Bereich der Metaanalyse schon seit mehreren Jahrzehnten. Neu ist, dass die Konsistenz der Ergebnisse in Untergruppen einer (Zulassungs-) Studie diskutiert wird und folglich die Überlegungen zur Power der Verfahren zum Nachweis inkonsistenter Ergebnisse eine größere Bedeutung bekommen. Zwar gibt es auch im Bereich der Metaanalyse Beispiele, dass Studien kombiniert wurden, die für sich selbst nicht in der Lage waren einen Wirksamkeitsnachweis zu erbringen, jedoch werden Studien üblicherweise so geplant, dass sie nach dem Wissen zum Zeitpunkt der Planung eine ausreichende Power haben, den interessierenden Effekt nachzuweisen. Die Untersuchung der Konsistenz des Behandlungseffekts in den Strata und Untergruppen einer Studie leidet dann zusätzlich daran, dass das Verhältnis der Strata durch die experimentelle Situation bestimmt wird (z. B. das Geschlechterverhältnis beim Herzinfarkt) und Studien eher selten so geplant werden, dass eine bestimmte wohldefinierte Evidenz in Untergruppen garantiert ist. In dieser Hinsicht ähnelt die Diskussion hier eher der Diskussion um die Untersuchung und Bewertung von Heterogenität bei mehrstufigen Versuchen in einem adaptiven Design (siehe z. B. die Diskussion in Friede und Henderson [22]).

Die Allgemeinen Methoden des IQWiG stellen korrekt fest, dass die Untersuchung der Heterogenität einer Patientenpopulation in Bezug auf die Konsistenz oder Ähnlichkeit des Therapieeffekts in der Literatur kontrovers diskutiert wird. In der Tat gibt es in der wissenschaftlichen Fachöffentlichkeit eine Diskussion darüber, unter welchen Bedingungen die Ergebnisse aus präspezifizierten und post hoc gefundenen Untergruppen als glaubwürdig eingestuft werden können. Hierzu verweisen die Allgemeinen Methoden des IQWiG auf Sun et al. [1] und vorausgegangene Arbeiten, die als eines der Kriterien für die Glaubwürdigkeit den Grad der Signifikanz des Heterogenitätstests bzw. des Interaktionsterms in einem statistischen Modell mit in die Bewertung einbeziehen. Diese Diskussion ist sicherlich auch der Tatsache geschuldet, dass in zunehmendem Maße Studien durchgeführt werden, in denen im Rahmen einer weltweiten Studie Wirksamkeit und Sicherheit eines Arzneimittels nachgewiesen werden sollen und es dann (in einer Studie an mehreren Tausend und gelegentlich bis zu 15.000 Patienten) schwierig ist, die Studienergebnisse ausschließlich auf der Basis eines Globalergebnisses zu akzeptieren oder zu verwerfen. In gleicher Weise verbietet sich die Frage nach einer Replikation des Studienergebnisses oder einer Metaanalyse zur Zusammenfassung von Subgruppen aus mehreren Studien.

Die Komplexität der Situation und die prinzipielle Schwierigkeit ein einheitliches Vorgehen zu definieren, mag der Tatsache geschuldet sein, dass mehrere ungünstige Faktoren zusammentreffen, die im Folgenden genannt werden. In allen Fällen

stellt sich die Frage, ob ein signifikanter Heterogenitätstest ein gutes Kriterium für die Glaubwürdigkeit eines Behandlungseffekts in einer Untergruppe darstellt und aus statistischer und erkenntnislogischer Sicht wohlbegründet ist.

Gemäß den Regeln der Arzneimittelentwicklung wird unter Umständen durch inkonsistente Studienergebnisse in Untergruppen ein globaler Wirksamkeitsnachweis gänzlich in Frage gestellt oder kann zu einer Einschränkung der Indikation führen [23]. In der Diskussion um Subgruppen im Rahmen der Nutzenbewertung wird aber genau das zum Kriterium für die Glaubwürdigkeit des Behandlungseffekts erhoben, was im Rahmen der Zulassung den prinzipiellen Wirksamkeitsnachweis infrage stellt, da die komplementäre Untergruppe dann eine deutlich geringere, keine Wirksamkeit oder sogar einen negativen Behandlungseffekt zeigt. Streng logisch betrachtet, sollte also derart große Heterogenität nach Zulassung in der indizierten Patienten- und Studienpopulation nicht vorkommen oder aber für den großen und respektive den kleinen (positiven) Therapieeffekt ist jeweils ein positives Nutzen-Risiko-Verhältnis durch die Zulassung bestätigt. Im hier betrachteten Modell wären dann inkongruente Entscheidungen nur dadurch zu erklären, dass Nichtunterlegenheit ausreichend für eine positive Zulassungsentscheidung ist.

Heterogenitätstests sind üblicherweise so konstruiert, dass ihre Nullhypothese die Annahme konsistenter Effekte darstellt, die auf der Basis beobachteter Daten abgelehnt werden soll. Geht es um die Bewertung der Konsistenz, so müsste eigentlich die Nullhypothese „bewiesen" werden. Folglich müssten, wenn überhaupt, große *p*-Werte als Nachweis der Konsistenz benutzt werden. Theoretisch wäre eine korrekte Darstellung als Äquivalenztest möglich. Dies würde jedoch erfordern, dass man für relevante Endpunkte zum Zeitpunkt der Studienplanung eine Aussage darüber treffen müsste, welche Unterschiede in den Therapieeffekten in Untergruppen als irrelevant eingestuft werden können. Vermutlich ist es nur in wenigen Fällen möglich entsprechende Annahmen zu rechtfertigen.

Der traditionelle Breslow-Day-Test auf Heterogenität hat, wie auch in dieser Arbeit dargestellt, unbefriedigende statistische Eigenschaften, wobei im Wesentlichen die Trennschärfe (Power) der Verfahren ein Problem ist. Weshalb dennoch oft signifikante Heterogenitätstests in der Bewertung der Studienergebnisse ignoriert werden, ist unter Berücksichtigung dieses Aspekts völlig unklar.

Nach wie vor ist das Bewusstsein innerhalb der wissenschaftlichen Gemeinschaft nicht ausgeprägt genug, dass der Wert eines Arzneimittels (oder allgemein einer medizinischen Therapie) nicht nur durch eine Betrachtung der Wirksamkeit alleine möglich ist, sondern zusätzlich die Sicherheit der Anwendung diskutiert werden muss und letztlich Nutzen (der die Bewertung der Wirksamkeit einschließt) und Risiken gegeneinander balanciert werden müssen.

Für die Nutzen-Risiko-Abwägung wäre die Vorschaltung eines Heterogenitätstests auf der Seite der Wirksamkeitsbewertung kontraproduktiv, da die Wahrnehmung von Inkonsistenzen behindert würde, die ja vielfach eher ein Schätzproblem sind und die tatsächliche Größe des Effekts eine wesentliche Rolle spielt. Obgleich also viele der in Sun et al. genannten Kriterien für die Glaubwürdigkeit von Ergebnissen aus Untergruppen relevante Aspekte der Diskussion beleuchten, fehlen die Überlegungen dazu, wie ein globales Nutzen-Risiko-Verhältnis in einer Untergruppe zu bewerten ist. Dieser Aspekt ist einer der wesentlichen Unterschiede zu den im Rahmen der jetzt als Draft verfügbaren Guideline zur Bewertung von Ergebnissen in Untergruppen einer randomisierten klinischen Studie formulierten Anforderungen: Sollen regulatorische Entscheidungen auf Untergruppen basieren, so ist auch zu prüfen, ob diese Untergruppen eine positive Nutzen-Risiko-Balance rechtfertigen oder doch wenigstens nicht infrage stellen [13].

Die Untersuchung der Konsistenz von Behandlungseffekten in den relevanten Untergruppen einer Studie ist ein wesentlicher Schritt in der Bewertung der Tragfähigkeit der Ergebnisse einer klinischen Studie der Phase III. Soll das Globalergebnis einer Studie Grundlage der Nutzenbewertung sein, sollten eigentlich dieselben Überlegungen auch in Bezug auf die Anwendbarkeit des Heterogenitätstests gelten.

Optimale Tests, die Signale für inkonsistente Studienergebnisse in Untergruppen der Studienpopulation anzeigen könnten und dabei eine hohe Sensitivität und Spezifität aufweisen würden, sind gegenwärtig nicht verfügbar. Empirisch begründete Regeln, wie in dieser Arbeit untersucht, sind ebenfalls nur in speziellen Situationen hilfreich. Häufig wird die Bewertung der Heterogenität mithilfe von I^2 als Mittel der Wahl angesehen, aber auch diesbezüglich wurde nachgewiesen, dass die Bewertung der Heterogenität zu falsch-positiven Befunden führen wird und deshalb nur eingeschränkt nutzbar ist [24]. Rücker et al. argumentieren dafür, direkt tau^2, den Varianzparameter im REM, zu schätzen, beantworten aber auch nicht die Frage, wann Studienergebnisse in einer Metaanalyse oder Therapieeffekte in Untergruppen als heterogen angesehen werden sollten. Diese sicherlich nicht befriedigende Tatsache kann, da es um die korrekte Bewertung von Nutzen und Risiken von Arzneimitteln und deren Anwendung geht, einer prinzipiellen Bewertung des Nutzen-Risiko-Profils in relevanten Untergruppen der Studienpopulation nicht entgegenstehen.

Falls optimale Tests nicht zur Verfügung stehen, muss im Zulassungskontext und im Kontext der Nutzenbewertung auf der Basis eines Globalergebnisses im Zweifelsfall einem Verfahren mit höherer Sensitivität der Vorzug vor Verfahren gegeben werden, die eine hohe Spezifität aufweisen. Heterogenität auf der Seite der Bewertung der Wirksamkeit ist ein Signal, dass eine weiterführende Diskussion um Nutzen und Risiken der Therapie erforderlich ist. So könnte ein bestimmtes Nebenwirkungsprofil eines Arzneimittels nicht mehr akzeptabel sein, wenn in einem frühen Stadium der Erkrankung nur ein reduzierter Therapieeffekt vorliegt.

Weiterentwicklungen der Methodik, die es erlauben würden, Heterogenität besser zu detektieren, sind dringend erforderlich. Erforderlich ist auch der korrekte Umgang mit Signalen, die in diesem

Prozess generiert werden: Selbst wenn bessere Tests zur Verfügung stehen würden, müsste es trotzdem möglich sein, ein Signal nach sorgfältiger inhaltlicher Prüfung der Rationale aus substanzwissenschaftlicher und klinisch-pharmakologischer Sicht als Zufallsbefund einzustufen. Nicht akzeptiert werden kann jedoch, dass ein auffälliger Heterogenitätstest ohne eingehende Prüfung a priori als Zufallsbefund eingestuft wird.

Zusammenfassend hoffen wir, in dieser Arbeit Argumente dafür geliefert zu haben, dass sich die Rolle des Heterogenitätstests weder im Bereich der Konsistenzprüfung der Ergebnisse von klinischen Studien noch in der Bemessung der Wertigkeit von Untergruppenergebnissen unterscheiden sollte. Insbesondere im letzten Fall ist es unplausibel, die Einschätzung des besonderen Werts eines Arzneimittels in einer Untergruppe vom (schlechten) Ergebnis desselben Arzneimittels in einer anderen Untergruppe einer klinischen Studie abhängig zu machen.

Vor dem Hintergrund einer ähnlichen Verantwortung und der Notwendigkeit, häufig auf der Basis gleicher Daten eine Entscheidung zu treffen, sehen wir einen großen Nutzen darin, Entscheidungsstrategien so weit möglich anzugleichen. Durch derartige Bemühungen wird dann auch genauer deutlich, an welchen Stellen Unterschiede zwischen Zulassungsentscheidung und Bewertung des Zusatznutzens bestehen und gegebenenfalls die Erhebung weiterer Daten erforderlich oder eine gemeinsame Diskussion der Studien gewinnbringend ist, die dann sowohl die Zulassung als auch die Bewertung eines Zusatznutzens begründen können.

Korrespondenzadresse

Prof. Dr. A. Koch
Institut für Biometrie
Medizinische Hochschule Hannover
Carl-Neuberg-Str. 1, 30625 Hannover
Koch.Armin@mh-hannover.de

Danksagung. Wir danken den Gutachterinnen und Gutachtern für die hilfreichen Kommentare, die die Diskussion um die Rolle der Heterogenitätsbewertung bereichert haben.

Einhaltung ethischer Richtlinien

Interessenkonflikt. Prof. Dr. A. Koch, A. Gonnemann und M. Kottas geben an, dass kein Interessenkonflikt besteht.

Dieser Beitrag beinhaltet keine Studien an Menschen oder Tieren.

Literatur

1. Sun X, Briel M, Walter SD, Guyatt GH (2010) Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ 340:850–854
2. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) (2013) Allgemeine Methoden 4.1. https://www.iqwig.de/download/IQWiG_Methoden_Version_4-1.pdf. Zugegriffen: 28. Okt. 2014
3. CAPRIE Steering Committee (1996) A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). Lancet 348:1329–1339
4. Bender R, Koch A, Skipka G et al (2010) No inconsistent trial assessments by NICE and IQWiG: different assessment goals may lead to different assessment results regarding subgroup analyses. J Clin Epidemiol 63:1305–1307
5. Hasford J, Bramlage P, Koch G et al (2010) Inconsistent trial assessments by the National Institute for Health and Clinical Excellence and IQWiG: standards for the performance and interpretation of subgroup analyses are needed. J Clin Epidemiol 63:1298–1304
6. Mok T S, Wu Y, Thongprasert S et al (2009) Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med 361:947–957
7. Mahaffey KW, Wojdyla DM, Carroll K et al (2011) Ticagrelor compared with clopidogrel by geographic region in the Platelet Inhibition and Patient Outcomes (PLATO) trial. Circulation 124:544–554
8. Committee for Proprietary Medicinal Products (2002) Points to consider on multiplicity issues in clinical trials. CPMP/EWP/908/99. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf. Zugegriffen: 28. Okt 2014
9. Sun X, Briel M, Busse JW et al (2012) Credibility of claims of subgroup effects in randomised controlled trials: systematic review. BMJ 344:1–9
10. Sun X, Heels-Ansdell D, Walter SD et al (2011) Is a subgroup claim believable? A user's guide to subgroup analyses in the surgical literature. J Bone Joint Surg Am 93:e8(1–9)
11. Sun X, Briel M, Busse JW et al (2011) The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. BMJ 342:1–8
12. Oxman AD, Guyatt GH (1992) A consumer's guide to subgroup analyses. Ann Intern Med 116:78–84
13. Committee for Medicinal Products for Human Use (2014) Guideline on the investigation of subgroups in confirmatory clinical trials – DRAFT. EMA/CHMP/539146/2013. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf. Zugegriffen: 28. Okt 2014
14. Chen J, Quan H, Binkowitz B et al (2010) Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. Pharm Stat 9:242–253
15. Quan H, Li M, Chen J et al (2010) Assessment of consistency of treatment effects in multiregional trials. Drug Inf J 617–632
16. Gail M, Simon R (1985) Testing for qualitative interactions between treatment effects and patient subsets. Biometrics 41:361–372
17. Piantadosi S, Gail MH (1993) A comparison of the power of two tests for qualitative interactions. Stat Med 12:1239–1248
18. Li J, Chan ISF (2006) Detecting qualitative interactions in clinical trials: an extension of range test. J Biopharm Stat 16:831–841
19. Bagheri Z, Ayatollahi SMT, Jafari P (2011) Comparison of three tests of homogeneity of odds ratios in multicenter trials with unequal sample sizes within and among centers. BMC Med Res Methodol 11:1–8
20. DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 7:177–188
21. Koch A, Framke T (2014) Reliably basing conclusions on subgroups of randomized clinical trials. J Biopharm Stat 24:42–57
22. Friede T, Henderson R (2009) Exploring changes in treatment effects across design stages in adaptive trials. Pharm Stat 8:62–72
23. Committee for Medicinal Products for Human Use (1998) ICH Topic E9: Statistical principles for clinical trials. CPMP/ICH/363/96. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf. Zugegriffen: 28. Okt 2014
24. Rücker G, Schwarzer G, Carpenter JR, Schumacher M (2008) Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol 8:79