Hans-Peter Dürr und Martin Scott

AMNOG und die Ermittlung eines Zusatznutzens:

Moderne Ergebnisvisualisierung anstelle von Mammut-Dossiers!

Die Nutzenbewertung neuer Arzneimittel erfolgt in der Regel auf der Grundlage einer Vielzahl von statistischen Analysen zu verschiedenen Endpunkten, Subgruppen, oder auch Patientenpopulationen. Dies erzeugt die Problematik des sogenannten *Multiplen Testens*, das sowohl pharmazeutische Unternehmer als auch Zulassungsbehörden mit einem statistischen Dilemma konfrontiert: Wie kann im Falle multiplen Testens die *Irrtumswahrscheinlichkeit* geeignet kontrolliert werden? Dieser Artikel beschreibt die Ursachen des Problems und zeigt auf, dass eine verbesserte Übereinkunft zur Visualisierung von Ergebnissen statistischer Tests beiden am Prozess beteiligten Parteien helfen würde: Einerseits dem pharmazeutischen Unternehmer beim Entwerfen der vielbeschworenen ‚Value Story' für sein Präparat, und andererseits den Zulassungsbehörden bei der Beurteilung derselben.

Zusammenfassung

Die Nutzenbewertung eines Arzneimittels basiert meist auf einer Vielzahl von statistischen Tests. Übermäßiges statistisches Testen kann jedoch Effekte postulieren, die nicht existieren. Dies betrifft nicht nur positive Effekte, wie z. B. die Wirksamkeit eines Medikamentes betreffend, sondern auch negative Effekte, wie z. B. Nebenwirkungen. Methoden, um diese Fehler zu korrigieren, existieren, sie führen jedoch – wenn konsequent durchgeführt – zum entgegengesetzten Problem: tatsächlich vorliegende Effekt werden verneint, auch hier Wirksamkeit und Nebenwirkungen betreffend. Die vom G-BA und dem IQWiG formulierten Vorgaben zur Zahl der Analysen und der Korrektur des sog. multiplen Testens sind stellenweise nur schwer zur Deckung zu bringen. Eine Lösung des Problems scheint nicht in Sicht, gleichzeitig könnte eine verfügbare Option besser genutzt werden: die Verwendung moderner Methoden zur Visualisierung von Ergebnissen.

Schlüsselwörter

AMNOG, Nutzenbewertung, Statistik, Multiples Testen

Problemstellung

>> Die Nutzenbewertung neuer Arzneimittel erfolgt in Deutschland durch den Gemeinsamen Bundesausschuss (G-BA), zusammen mit dem Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), welches die methodischen und statistischen Vorgaben für den Bewertungsprozess festlegt. Der Zusatznutzen eines neuen Medikaments wird dabei im Rahmen von klinischen Studien oder Register-Studien untersucht, durch statistische Auswertungen der Studiendaten quantifiziert, und schließlich eingeteilt in die Kategorien geringer, beträchtlicher oder erheblicher Zusatznutzen.

Die statistischen Auswertungen führen dabei zu einem Problem, das seit den Anfängen der Statistik bekannt ist: Wer viel sucht, der wird mit höherer Wahrscheinlichkeit etwas finden. Dieses Problem des *Multiplen Testens* ergibt sich auch für das Auffinden von Behandlungseffekten hinsichtlich der Wirksamkeit eines Medikaments. Durch übermäßiges statistisches Testen kann man zum Beispiel zu dem überraschenden Ergebnis kommen, dass Aspirin besonders wirksam sei für Menschen, die im Sternzeichen Steinbock geboren wurden, völlig unwirksam hingegen für diejenigen, die im Sternzeichen Waage oder Zwilling geboren wurden (Peto et al. 1995). Könnte das ein Irrtum sein, der auf einem methodischen Problem beruht? Ja – natürlich ist dies der Fall.

Ein zentraler Begriff in diesem Kontext ist die sogenannte *Irrtumswahrscheinlichkeit* α. Sie gibt an, mit welcher Wahrscheinlichkeit ein Effekt, der in Wahrheit nicht vorliegt, als statistisch signifikant dargestellt wird und damit einer falsch-positiven Schlussfolgerung gleichkommt. ‚Statistisch signifikant' bedeutet dabei, dass es unwahrscheinlich ist, diesen Effekt zufällig zu finden, wenn es ihn in Wahrheit nicht gibt. Dabei wird ein Ereignis in der Regel dann als ‚unwahrscheinlich' bezeichnet, wenn es – unter Annahme von Zufälligkeit (der Nullhypothese) – in weniger als 5% der Fälle auftritt ($\alpha<5\%$). Liegt eine Wirksamkeit in Wahrheit nicht vor, und die statistische Analyse postuliert dennoch eine solche Wirksamkeit, dann geht dieser Irrtum auf die Irrtumswahrscheinlichkeit zurück. Der statistische Ansatz fällt gemäß der Irrtumswahrscheinlichkeit in 5% der Fälle einem systematischen Irrtum zum Opfer.

Die Konsequenz des Multiplen Testens ist: Führen wir 20 statistische Tests durch, und jeder Test begeht mit einer Wahrscheinlichkeit von 5% einen Fehler, dann werden wir durchschnittlich bei einem von 20 Testergebnissen rein zufällig eine Wirksamkeit postulieren (mit Wahrscheinlichkeit von 20 x 5% = 100% !). Dies führt zum Kern des Problems: Für die Nutzenbewertung eines Medikamentes werden in der Regel nicht nur 20 statistische Tests durchgeführt, sondern Hunderte, nicht selten Tausende. Das macht es wahrscheinlich, eine Wirksamkeit zu postulieren, die in Wirklichkeit nicht vorliegt. Verschärft wird diese Problematik bei der medizinischen Nutzenbewertung durch die vorgeschriebenen Interaktions- und Subgruppenanalysen, die die Anzahl der durchzuführenden statistischen Analysen immens in die Höhe treiben – und damit auch die Wahrscheinlichkeit eines Irrtums.

Häufig zu untersuchende Subgruppenmerkmale sind z. B. Alter, Geschlecht, Schweregrad oder Stadium der Erkrankung, vorhergehende Behandlungen, etc. Relevant für die Patientengesundheit ist die Frage, ob die Wirksamkeit eines Medikaments vom Subgruppenmerkmal abhängt. Am Beispiel des Subgruppenmerkmals ‚Geschlecht' würde interessieren, ob das Medikament bei Männern besser wirkt als bei Frauen, oder umgekehrt. Ist das der Fall, liegt eine sogenannte Wechselwirkung vor, auch Interaktion genannt. Die von G-BA und IQWiG aufgestellten Vorgaben verlangen, Subgruppenmerkmale ausführlich zu berücksichtigen. Gleichwohl gibt das IQWiG zu bedenken, *„dass die Ergebnisse irgendeiner Subgruppe statistische Signifikanz erreichen, obwohl es sich in Wahrheit um ein zufälliges Ergebnis handelt"* (IQWiG 2017).

Lösungsansätze

Es gibt mehrere Methoden, um die Problematik des multiplen Testens zu kontrollieren (Bender et al. 2007). Allen diesen Methoden ist die Absenkung des Signifikanzniveaus gemeinsam, was das Auffinden von erwünschten Effekten schwieriger macht. In übermäßiger Zahl durchgeführt ruinieren statistische Tests also die Wahrscheinlichkeit, die Wirksamkeit eines Medikaments nachweisen zu können, oder anders gesagt: Gerade der Versuch, eine Wirksamkeit nachzuweisen, verschlechtert die Chance, diesen Nachweis statistisch erbringen zu können.

In der einfachsten Methode, der sog. Bonferroni-Korrektur, wird das Signifikanz-Niveau dadurch abgesenkt, dass die Irrtumswahrscheinlichkeit α dividiert wird durch die Anzahl der durchgeführten, statistischen Tests – je mehr getestet wird, desto höher wird die Anforderung für die Aussage ‚Test bestanden'. Bei 100 statistischen Tests müsste das Signifikanzniveau von α=5% auf α=0.05% abgesenkt werden. Die Wirksamkeit eines Medikamentes würde nur noch dann ‚statistisch signifikant' werden, wenn der Effekt sehr stark oder die Zahl der untersuchten Patienten sehr hoch ist.

Beim Thema Nebenwirkungen tritt das umgekehrte Problem auf: Das abgesenkte Signifikanzniveau setzt die Nachweisschwelle für Nebenwirkungen so tief, dass in der Praxis viele unerwünschte Nebenwirkungen, die in Wirklichkeit auftreten, als ‚statistisch nicht signifikant' bezeichnet würden. Die wissenschaftlich korrekte und konsequente Umsetzung von Bonferroni-Methoden führt zu einer über-konservativen Ablehnung von Effekten, sowohl die Wirksamkeit eines Medikaments als auch das Risiko unerwünschter Nebenwirkungen betreffend.

Bereits 1998 umschrieb Feinstein die Problematik als eine ‚clinicostatistical tragedy' und forderte ein ‚weg von multiplem Testen, zurück zu klinischer Relevanz' (Feinstein 1998). Er ging insbesondere auf das Problem ein, dass die Zahl der statistischen Analysen durch Interaktions- und Subgruppenanalysen stark steigt und letztlich Fehlbefunde durch ‚data dredging' geschaffen werden, also einem eher ungezielten ‚Ausquetschen' von Studiendaten. Die von ihm beschriebene Problematik ist aktueller denn je, auch wenn sie nach der Jahrtausendwende durch positiv besetzte Begriffe wie ‚data mining' oder ‚big data' salonfähiger geworden ist.

Gegenwärtig gibt es keine allgemein anerkannte Vorgehensweise, die wissenschaftlich korrekt und gleichermaßen praktikabel ist; in dem Sinne, dass sie einerseits mit hinreichender Wahrscheinlichkeit die Wirksamkeit eines Medikaments entdeckt und andererseits das Risiko durch Nebenwirkungen angemessen ausschließt. Die im Februar 2019 veröffentlichte Dokumentvorlage des G-BA zur Erstellung von Dossiers lässt jedoch den Trend erkennen, die Visualisierung von Ergebnissen statistischer Tests mehr in den Vordergrund zu stellen (G-BA 2019), beispielsweise in Form von Forest-Plots und Überlebenszeitkurven.

Ausblick

Die derzeitige Praxis der Nutzenbewertung von Medikamenten leidet unter einem methodischen Interessenkonflikt. Der Versuch, möglichst viele Detailfragen testen und beantworten zu wollen, kollidiert mit der statistischen Problematik des multiplen Testens. Statistische Methoden zur Korrektur des multiplen Testens (z.B. die oben erwähnte Bonferroni-Korrektur) können häufig nicht konsequent eingesetzt werden, weil dies dazu führen würde, dass sowohl Wirksamkeit als auch Schädlichkeit eines Medikaments nicht ausreichend sensitiv entdeckt werden. Die von Feinstein geprägte Formulierung der ‚clinicostatistical

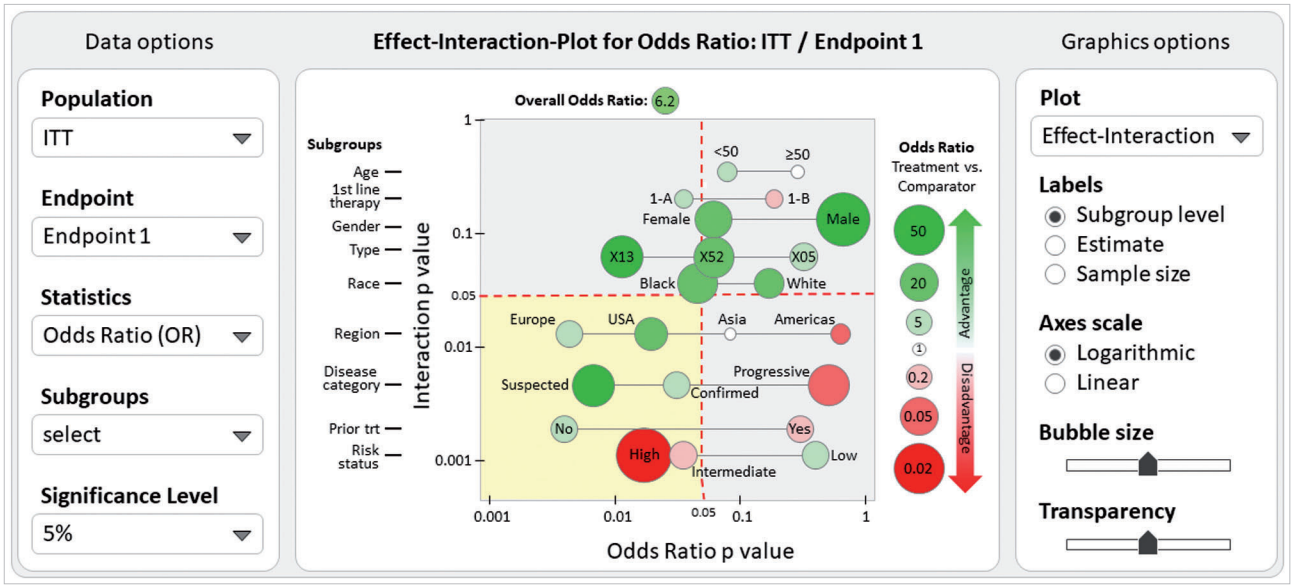

Abbildung 1: Effekt-Interaktions-Graph zur Darstellung von Interaktions- und Subgruppenanalysen im Rahmen von Nutzenbewertungen (https://de.numerus.com/IEBfree). Der Behandlungseffekt wird in diesem Beispiel durch das Odds Ratio (OR) repräsentiert und für einen Endpunkt und die ITT-Population dargestellt. Y-Achse: p-Wert für die Interaktion des Behandlungseffekts mit dem Subgruppenmerkmal (Age, Gender, etc.). X-Achse: p-Wert des Behandlungseffekts für eine Ausprägung des Subgruppenmerkmals (z. B. Age: < 50 und ≥ 50). Ausprägungen innerhalb einer Subgruppe sind durch horizontale Linien verbunden. Grün: günstiger Behandlungseffekt, Rot: ungünstiger Behandlungseffekt. Effektstärken (hier: OR) werden durch einen proportionalen Kreisdurchmesser der ‚bubbles' visualisiert (s. Legende rechts). Der Gesamt-Effekt ist am oberen Rand der Grafik dargestellt (OR=6.2). Der gelb eingefärbte Quadrant enthält die statistisch signifikanten Effekte (hier: α= 0.05). Daten- und Grafikoptionen können im linken und rechten Panel hinsichtlich der spezifischen Fragestellung einfach und schnell angepasst werden. Das Datenbeispiel zeigt eine Situation, in welcher der größte Nachteil des Medikaments in der Patienten-Subgruppe mit Risk-Status ‚High' festzustellen ist. Quelle: Numerus

tragedy' beschreibt das Ausmaß dieses Dilemmas auf drastische Weise. Es scheint eine Ära betreten worden zu sein, in der übermäßiges statistisches Testen in den Hintergrund tritt und durch eine interaktiv-explorative, ausgewogene Gesamtschau der Ergebnisse abgelöst wird.

Die theoretische Lösung für das dargestellte Problem ist zunächst klar: Strukturell liegt ein Rating- oder Scoring-Problem vor. Die derzeitige Einteilung des Zusatznutzens in die Kategorien gering, beträchtlich oder erheblich ist das Ergebnis eines solchen Rating-Verfahrens. Ein eindeutiger Rating-Algorithmus kann jedoch nicht zugrunde gelegt werden, weil Hoch-Dimensionalität vorliegt: Eine Nutzenbewertung erfolgt regelmäßig in den Dimensionen Population(en), Endpunkt(e), Subgruppe(n), und Auswertungsmethode(n). Jede dieser Dimensionen enthält mehrere, ja Dutzende von Ausprägungen; die kombinatorische Matrix erreicht dann schnell Hunderte oder gar Tausende von Einzelergebnissen. Jedes dieser Einzelergebnisse müsste für sich gewichtet werden, weil verschiedene Endpunkte – z. B. Mortalität und Lebensqualität – regelmäßig nicht dieselbe Patientenrelevanz haben. In der Summe bräuchten solche Verfahren also Tausende von Expertenmeinungen, und das für jede Erkrankung separat. Man erkennt, dass die Grenzen von Machbarkeit schnell überschritten werden. Was kann hier helfen?

Wenn die praktische Lösbarkeit des Problems derart erschwert ist, führt dies zurück zur oben beschriebenen Strategie einer geeigneten Darstellung und Visualisierung der Ergebnisse. Ein mehrere Tausend Seiten umfassendes Dossier zur Nutzenbewertung eines Medikaments kann behaupten, ‚umfassend' zu sein, jedoch nicht, ‚Übersichtlichkeit' anzubieten. Wünschenswert wäre für alle Beteiligten, dass statistische Resultate in einer standardisierten und übersichtlichen Form zur Beurteilung angeboten werden, wie in der Abbildung oben gezeigt. Hiervon würden alle Beteiligten profitieren: Der pharmazeutische Unternehmer könnte die Chancen und Risiken seines Medikamentes einfacher und sicherer erkunden und darstellen. Die Zulassungsbehörden könnten schneller und verlässlicher zu einer Gesamteinschätzung des Nutzens der Behandlung kommen.

Vielleicht fasst die nachfolgende Formulierung, die zurzeit häufig Anwendung findet, die Problematik und das Verbesserungspotenzial geeignet zusammen: *Dialog und Konsensfindung auf Augenhöhe* – Dies könnte im Falle der Nutzenbewertung von Arzneimitteln bedeuten: Durch eine geeignete Visualisierung der Ergebnisse statistischer Analysen. <<

Literatur

Bender R, Lange S and Ziegler A (2007) [Multiple testing]. Dtsch Med Wochenschr 132 Suppl 1:e26-29.

Gemeinsamer Bundesausschuss (2019) Anlage II.6: Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen. Dokumentvorlage, Version vom 21.02.2019.

Feinstein AR (1998) The problem of cogent subgroups: a clinicostatistical tragedy. J Clin Epidemiol 51:297-299.

IQWiG (2017) Allgemeine Methoden: Version 5.0, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln.

Peto R, Collins R and Gray R (1995) Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 48:23-40.

PD Dr. Hans-Peter Dürr
ist Gesundheitswissenschaftler und hat sich an der Universität Tübingen im Fach Epidemiologie habilitiert. Neben seinen Forschungsthemen im Bereich der mathematischen und statistischen Modellierung in der Medizin leitet er bei der Firma Numerus die Abteilung Epidemiologie und wissenschaftliche Methodik und ist zuständig für epidemiologische Fragestellungen und statistische Analysen in den Bereichen HTA/Nutzenbewertung, Modellierung und Simulation.

Kontakt: hans-peter.duerr@numerus.com

Martin Scott
ist Statistiker und Geschäftsführer der Numerus GmbH Deutschland und arbeitet auf dem Gebiet der Arzneimittel- und Medizinproduktforschung in verschiedenen therapeutischen Bereichen und Studientypen, einschließlich Phase IV, Registrierungsstudien, HEOR und HTA. Er arbeitete zuvor in mehreren globalen Pharmaunternehmen und klinischen Forschungsorganisationen und berät gegenwärtig Market Access Teams unter besonderer Berücksichtigung der Vorgaben des AMNOG in Deutschland.

Kontakt: martin.scott@numerus.com