IQWiG Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Tiotropiumbromid bei COPD

Dokumentation und Würdigung der Anhörung zum Berichtsplan

Auftrag A05-18
Version 1.0
Stand: 02.08.2010

Impressum

Herausgeber:
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Thema:
Tiotropiumbromid bei COPD

Auftraggeber:
Gemeinsamer Bundesausschuss

Datum des Auftrags:
22.02.2005

Interne Auftragsnummer:
A05-18

Anschrift des Herausgebers:
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Dillenburger Str. 27
51105 Köln

Tel.: +49 221 35685-0
Fax: +49 221 35685-1
Berichte@iqwig.de
www.iqwig.de

Inhaltsverzeichnis

Seite

Abkürzungsverzeichnis

Abkürzung	**Bedeutung**
6MWT	6-Minuten-Gehtest (6-Minute Walk Test)
ADAS-cog	Alzheimer's Disease Assessment Scale, cognitive subscale
AMNOG	Arzneimittelmarktneuordnungsgesetz
ATS/ERS	American Thoracic Society/European Respiratory Society
CAT	COPD Assessment Test
CONSORT	Consolidated Standards of Reporting Trials
COPD	chronisch obstruktive Lungenerkrankung (chronic obstructive pulmonary disease)
CPET	Cardiopulmonary Exercise Testing
EMA (ehemals EMEA)	European Medicines Agency
FDA	Food and Drug Administration
FEV_1	Einsekundenkapazität, forciertes exspiratorisches Volumen in einer Sekunde (forced expiratory volume in 1 second)
FVC	forcierte Vitalkapazität (forced vital capacity)
GKV	gesetzliche Krankenversicherung
GRADE	Grading of Recommendations Assessment, Development and Evaluation
IC	inspiratorische Kapazität (inspiratory capacity)
ICS	inhalative Glucocorticosteroide
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
ITT	Intention-to-Treat
LABA	langwirksame Beta-2-Rezeptoragonisten
LCA	latent class analysis
LOCF	last observation carried forward
GOLD	Global Initiative for Chronic Obstructive Lung Disease
MCID	minimal clinically important difference
MID	minimal important difference
NNT	number needed to treat
RCT	randomisierte kontrollierte Studie (randomised controlled trial)
TDI	Transition Dyspnea Index
SMD	standardisierte Mittelwertdifferenz

Abkürzung	Bedeutung
SGRQ	St. George’s Respiratory Questionnaire

1 Dokumentation der Anhörung

Am 19.04.2010 wurde der vorläufige Berichtsplan in der Version 1.0 vom 12.04.2010 veröffentlicht und zur Anhörung gestellt. Bis zum 18.05.2010 konnten schriftliche Stellungnahmen eingereicht werden. Insgesamt wurden 5 Stellungnahmen form- und fristgerecht abgegeben. Diese Stellungnahmen sind im Anhang abgebildet.

Unklare Aspekte in den schriftlichen Stellungnahmen wurden in einer wissenschaftlichen Erörterung am 16.06.2010 im IQWiG diskutiert. Das Wortprotokoll der Erörterung befindet sich in Kapitel 4.

Eine Würdigung der in der Anhörung vorgebrachten Aspekte befindet sich in Kapitel 2. Im überarbeiteten Berichtsplan sind darüber hinaus Änderungen, die sich durch die Anhörung ergeben haben, zusammenfassend dargestellt. Der überarbeitete Berichtsplan ist auf der Website des IQWiG unter www.iqwig.de veröffentlicht.

2 Würdigung der Anhörung

Die im Rahmen der Anhörung vorgebrachten Aspekte wurden hinsichtlich valider wissenschaftlicher Argumente für eine Änderung des Berichtsplans überprüft. Die wesentlichen Argumente werden im Folgenden diskutiert. Neben projektspezifischen wissenschaftlichen Aspekten wurden auch übergeordnete Punkte, z. B. zu Verfahrensschritten gemäß rechtlichen Vorgaben für das Institut, angesprochen. Auf solche Punkte wird im Rahmen dieser projektspezifischen Würdigung der Anhörung nicht weiter eingegangen.

2.1 Definition der der Bewertung zugrunde liegenden Zielpopulation

Die Formulierungen des Berichtsplans zur Zielpopulation der Nutzenbewertung wurden von Stellungnehmenden ohne Begründung als nicht „hinreichend explizite Angaben“ bezeichnet und daher wurden klarere Spezifikationen von Patienten mit chronisch obstruktiver Lungenerkrankung (COPD) in Abgrenzung zu Asthmapatienten gefordert.

Die im Berichtsplan als Einschlusskriterium formulierte Diagnose anhand von Kriterien anerkannter Leitlinien bei gleichzeitigem Ausschluss von Asthmapatienten entspricht in ihrer Detailtiefe den Vorgaben der Zulassungsbehörde [1]. Dies ist auch im Rahmen dieses Berichtsplans als zweckmäßig zu betrachten, da die Vorgaben deutscher und internationaler Leitlinien [2-5] eine klare Abgrenzung der beiden Diagnosen erlauben.

Zwischen einzelnen Leitlinien bestehen geringfügige Unterschiede z. B. im geforderten Wert der Reversibilität der Bronchokonstriktion zur Abgrenzung der COPD gegenüber Asthma. So betrachtet die ATS/ERS-Leitlinie[1] eine Bronchokonstriktion, die größtenteils („largely“) reversibel ist, als Hinweis auf eine Asthmadiagnose [3]. 2 deutsche Leitlinien geben für die Diagnose der COPD einen Wert von weniger als 15 % FEV_1-Erhöhung[2] an [2,5], die GOLD-Leitlinie[3] einen Wert von weniger als 12 % [4].

Deshalb ist es nicht zielführend, konkrete Angaben zur Differenzialdiagnose im Berichtsplan aufzuführen. Entscheidend wird bei der Auswahl der Studien sein, ob die Operationalisierung der Patienteneinschlusskriterien eindeutig und mit Verweis auf eine anerkannte Leitlinie begründbar ist. Sollten die Studienpopulationen bezüglich der Diagnosekriterien im Rahmen der Leitlinien voneinander abweichen, kann gegebenenfalls eine Effektmodifikation durch die Einschlusskriterien untersucht werden.

Daher ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans.

[1] American Thoracic Society/European Respiratory Society
[2] Erhöhung der Einsekundenkapazität
[3] Global Initiative for Chronic Obstructive Lung Disease

2.2 Definition von Intervention und Vergleichsbehandlung

In einer Stellungnahme wurde angemerkt, dass unklar sei, welche Therapieoptionen konkret in die Nutzenbewertung eingehen sollen.

In Abschnitt 4.1.2 des Berichtsplans finden sich die Angaben, dass Tiotropiumbromid verglichen werden soll mit Placebo bzw. mit einer anderen in Deutschland verfügbaren medikamentösen Therapieoption, jeweils als inhalative Dauertherapie der COPD. Der Berichtsplan führt zudem aus, dass die Anwendung der Prüf- und Vergleichsinterventionen in den Studien im Rahmen des für Deutschland gültigen Zulassungsstatus erfolgen muss.

Die Bewertung des IQWiG bewegt sich also innerhalb des Zulassungsstatus der zu prüfenden Arzneimittel. Das gilt auch für Studien, in denen gegebenenfalls Kombinationsbehandlungen untersucht wurden. Darüber hinaus müssen die zu prüfenden Interventionen in Deutschland verfügbar sein. Zulassungsstatus und Verfügbarkeit setzen einen eindeutigen Rahmen der zu bewertenden Prüf- und Vergleichsinterventionen. Weitere Einschränkungen werden im Berichtsplan nicht vorgenommen. Die zu bewertenden Interventionen sind damit klar definiert.

Zudem wurde von einem weiteren Stellungnehmenden angeregt, auch Studien auszuwerten, „in denen der Zusatznutzen von Tiotropium im Rahmen einer Rehabilitation dargestellt wird“.

Bei der in diesem Zusammenhang als Beispiel genannten Studie [6] handelt es sich um einen Vergleich von Tiotropiumbromid und Placebo, bei dem beide Gruppen zusätzlich einer Rehabilitationsmaßnahme unterzogen wurden. Dieses Szenario stellt laut Berichtsplan kein Kriterium für einen Ausschluss aus der Nutzenbewertung dar. Entsprechen also alle anderen Eigenschaften einer solchen Untersuchung den im Berichtsplan definierten Einschlusskriterien, werden Studien mit dieser Konstellation in die Nutzenbewertung einfließen.

Zusammenfassend ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans.

2.3 Endpunkte

In mehreren Stellungnahmen wurde Bezug auf die im Berichtsplan formulierten patientenrelevanten Endpunkte genommen. Die verschiedenen angesprochenen Aspekte zu den einzelnen Endpunkten werden nachfolgend diskutiert.

2.3.1 Angabe einer zu bewertenden Effektrichtung

Eine Stellungnahme regte an, den Endpunkt gesundheitsbezogene Lebensqualität um die Formulierung „Besserung bzw. Erhalt der“ zu ergänzen, wie dies in früheren IQWiG-Berichten

der Fall war. Dies wurde damit begründet, dass auch ein Erhalt dieser Größe bewertet werden sollte.

In früheren Berichtsplänen hat das IQWiG Zielgrößen formuliert, die patientenrelevanten Therapiezielen zuzuordnen waren. Mit Bezug auf die Therapieziele wurde deshalb von einer Besserung bzw. einem Erhalt der gesundheitsbezogenen Lebensqualität gesprochen.

In den aktuellen Berichtsplänen verzichtet das IQWiG auf den Bezug zu Therapiezielen und benennt direkt patientenrelevante Endpunkte. Die Benennung einer (gewünschten) Effektrichtung ist dabei, analog beispielsweise zur Benennung von Endpunkten in klinischen Studien, nicht sinnvoll. Die tatsächlichen Effektgrößen und -richtungen ergeben sich aus dem Vergleich der Werte für die Testintervention mit denen der Vergleichsbehandlung und werden anschließend im Kontext der Erkrankung sowie der Vergleichsbehandlung und anderer Studiencharakteristika interpretiert. Damit wird auch der Situation Rechnung getragen, dass bei einer progredienten Erkrankung bereits ein Erhalt der Lebensqualität ein Nutzen sein kann.

Daher ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans.

2.3.2 Angabe konkreter Messinstrumente

In einer Stellungnahme wurde angeregt, zu den patientenrelevanten Endpunkten bereits im Berichtsplan einzelne Messinstrumente beziehungsweise Operationalisierungen aufzunehmen.

Im Gegensatz zu einer Studie mit prospektiv geplanter Datenerhebung ist zum Zeitpunkt der Planung einer systematischen Übersicht die Datenstruktur der verfügbaren Daten im Sinne einer Zielgrößendefinition (z. B. hinsichtlich der verwendeten Messinstrumente beziehungsweise Operationalisierungen) für alle einzuschließenden Studien nicht bekannt. Der Versuch einer Aufzählung aller möglichen oder wahrscheinlich verwendeten Messinstrumente und Operationalisierungen von Endpunkten in den zu bewertenden Studien ist deshalb nicht zielführend. Es ergibt sich vielmehr die Notwendigkeit, mit den in der Recherche identifizierten Datenstrukturen umzugehen.

Die Entscheidung über den Einschluss von Messinstrumenten für die im Berichtsplan präspezifizierten Endpunkte erfolgt auf Basis der Methodik des IQWiG [7]. So müssen die Messinstrumente z. B. patientenrelevante Konstrukte abbilden. Das Methodenpapier beschreibt auch Anforderungen an patientenberichtete Endpunkte, die z. B. zur Untersuchung von Nutzendimensionen wie der gesundheitsbezogenen Lebensqualität oder zur Messung von Symptomen zum Einsatz kommen können. Die Definition solcher Kriterien für den Einschluss von Messinstrumenten wird der Durchführung einer systematischen Übersicht eher gerecht als der Versuch, alle möglicherweise zu bewertenden Instrumente bereits im Vorfeld zu definieren.

Daher wird keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans gesehen.

2.3.3 Verbrauch an Bedarfsmedikation

In einer Stellungnahme wurde die Auffassung vertreten, dass der Verbrauch an Bedarfsmedikation bei der Bewertung Berücksichtigung finden sollte, da er den Grad und / oder die Häufigkeit von auftretenden Symptomen für den Patienten reflektiere.

Die Stellungnehmenden benennen also den Verbrauch an Bedarfsmedikation als Surrogat für die COPD-Symptomatik. Damit stellt sich die Frage, ob der Verbrauch an Bedarfsmedikation ein valides Surrogat ist, ob es also eindeutige Belege aus Interventionsstudien gibt, die einen plausiblen, starken, konsistenten und gleichgerichteten Zusammenhang zwischen der Änderung des Surrogats (Bedarfsmedikation) und der Änderung des patientenrelevanten Endpunkts (COPD-Symptome) zeigen [7]. Die Stellungnehmenden haben dazu keine Literatur eingereicht.

In der Studie Casaburi 2005 [6] wurden Patienten mit COPD zusätzlich zu einer Rehabilitationsmaßnahme mit Tiotropiumbromid oder Placebo behandelt (siehe dazu auch Abschnitt 2.2). In dieser Studie war der Verbrauch an Albuterol als Bedarfsmedikation bereits nach einer Woche unter Tiotropiumbromid niedriger als unter Placebo. Dieser Unterschied blieb während der ganzen Studie relativ konstant. Die Gruppendifferenzen im Bereich der Dyspnoe (TDI[4]) und der gesundheitsbezogenen Lebensqualität (SGRQ[5]) variierten dagegen im Verlauf der Studie. Insbesondere spiegelte sich eine erneute Zunahme der Dyspnoe in der Placebogruppe gegen Ende der Studien nicht im Verbrauch an Bedarfsmedikation in der Placebogruppe wider. Die Studie zeigt also diskrepante Effekte für Bedarfsmedikation und Symptomatik. Dieses Ergebnis spricht gegen die Validität des Surrogats Bedarfsmedikation..

Auch die Leitlinie der US-amerikanischen und der europäischen Zulassungsbehörde sowie eine ATS/ERS-Task-Force zu „Outcomes for COPD pharmacological trials“ benennen den Verbrauch an Bedarfsmedikation nicht als Endpunkt für COPD-Studien [1,8,9].

In der Diskussion dieses Punktes in der mündlichen Erörterung wurden keine weiteren Belege dafür vorgebracht, dass es sich bei dieser Größe um ein valides Surrogat eines patientenrelevanten Endpunktes handelt. Es wurde lediglich diskutiert, dass dieser Parameter eventuell als Effektmodifikator wirken könnte. Von einigen Stellungnehmenden wurde darauf hingewiesen, dass der Stellenwert des Verbrauchs an Bedarfsmedikation bei der COPD geringer sei als bei Asthma. Die unterschiedliche Relevanz dieses Parameters zeige sich u. a. darin, dass sich Instrumente zur schnellen Beurteilung in der klinischen Praxis in diesem Punkt unterschieden. So erhebe der Asthma-Kontrolltest den Verbrauch an Bedarfs-medikation, während dies beim COPD Assessment Test (CAT) nicht der Fall sei [10,11].

Daher wird keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans gesehen.

[4] Transition Dyspnea Index
[5] St. George’s Respiratory Questionnaire

2.3.4 Körperliche Belastbarkeit

Der Berichtsplan benennt die körperliche Belastbarkeit als patientenrelevanten Endpunkt für die Bewertung von Tiotropiumbromid. Geeignete Maße zur Beschreibung der körperlichen Belastbarkeit sind Symptome unter Belastung und Parameter der Leistungsfähigkeit. Letztere lassen sich z. B. erfassen durch die mögliche Belastungsdauer oder die Gehstrecke beim – von einer Stellungnahme als Endpunkt vorgeschlagenen – 6-Minuten-Gehtest (6-Minute Walk Test [6MWT]) [9,12,13].

Darüber hinaus wurde in einer Stellungnahme der kardiopulmonale Belastungstest (Cardiopulmonary Exercise Testing [CPET]) als wesentliche Methode zur Untersuchung körperlicher Belastbarkeit bei der COPD genannt [14].

Zusätzlich wurde Literatur zur Interpretation der Effektgrößen der beiden Skalen erwähnt [12,15]. Diese wird in der Nutzenbewertung Berücksichtigung finden.

Beim CPET werden die Patienten mittels eines Fahrradergometers oder eines Laufbandes körperlicher Belastung ausgesetzt. Es können neben der maximalen Leistung und Ausdauer der Patienten auch der Gasaustausch in der Lunge sowie kardiale und ventilatorische Parameter untersucht und eine Blutgasanalyse durchgeführt werden. Das in der Stellungnahme zitierte Statement der American Thoracic Society/American College of Chest Physicians zum kardiopulmonalen Belastungstest räumt aber Forschungsbedarf ein in der Frage des Zusammenhangs zwischen den erhobenen physiologischen und den patientenberichteten Werten [14].

Daher wurde in der mündlichen Erörterung im Rahmen der Anhörung des vorläufigen Berichtsplans die Frage gestellt, welche der beim CPET erhobenen Messgrößen nach Ansicht der Stellungnehmenden in die Nutzenbewertung einfließen sollten.

Von den Stellungnehmenden wurde klargestellt, dass die Hauptintention der Stellungnahme gewesen sei, im Berichtsplan konkrete Messinstrumente zu den einzelnen Endpunkten zu nennen. Angaben dazu, welche spezifischen Messgrößen der CPET in die Nutzenbewertung einfließen sollten, wurden nicht gemacht.

Wie bereits in Abschnitt 2.3.2 ausgeführt, ist eine Aufzählung aller möglichen oder wahrscheinlich verwendeten Messinstrumente und Operationalisierungen von Endpunkten im Berichtsplan nicht zielführend. Es wird daher keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans gesehen.

2.3.5 Lebensqualität

Nach Auffassung einer Stellungnahme sollte der Begriff „Lebensqualität“ durch „Gesundheitsstatus“ ergänzt werden, denn die auf dem Gebiet der COPD gebräuchlichen Fragebogeninstrumente (wie z. B. der St. George’s Respiratory Questionnaire – SGRQ)

würden ein quantitatives Maß für den Gesundheitsstatus darstellen, ermittelten aber nicht die Lebensqualität.

In der angegebenen Literatur äußert sich Jones zwar so, wie dies die Stellungnehmenden zitierten, d. h. er bezeichnet den SGRQ als Gesundheitsstatus-Fragebogen und tritt für eine klare Differenzierung zwischen Gesundheitsstatus („health status“) und Lebensqualität („quality of life“) ein. Er nimmt jedoch keine klare Abgrenzung der beiden Konzepte vor [16]. So bezeichnet der gleiche Autor in seiner Publikation zur Entwicklung des SGRQ den Fragebogen als Instrument zur Messung von Lebensqualität („questionnaire for measuring impaired health and perceived well-being [’quality of life’])“ [17]. Er benennt die 3 Domänen des Fragebogens mit „Symptome“, „Aktivität“ und „Belastung (impact)“ und führt aus, dass „Belastung“ Aspekte der sozialen Funktionsfähigkeit und von psychologischen Beeinträchtigungen abbildet. Dabei handelt es sich um Domänen der gesundheitsbezogenen Lebensqualität.

Generell sind die Begriffe „Gesundheitsstatus“ und „gesundheitsbezogene Lebensqualität“ in der Literatur nicht scharf abgegrenzt. Chassany schreibt dazu, dass es keine endgültig konsentierte Definition von gesundheitsbezogener Lebensqualität gibt. Bei der Verwendung diverser Begrifflichkeiten („health status, well-being, quality of life, health-related quality of life“) würden die Autoren das gleiche Konzept mit unterschiedlichen Worten zum Ausdruck bringen [18, S. 212]. Fayers stellt fest, dass für viele Jahre einige Fragebögen den Schwerpunkt auf „health status“ und „self-reported health“ gelegt hätten bei gleichzeitig beträchtlicher Überschneidung mit dem Konzept Lebensqualität. So verwendet er in seinem Buch einheitlich den etablierten Begriff „quality of life“ [19, S. 3].

Da die Konzepte also nicht trennscharf voneinander abzugrenzen sind, wird keine Notwendigkeit gesehen, den Begriff „Gesundheitsstatus“ in den Berichtsplan aufzunehmen. Entsprechend merkt eine andere Stellungnahme an, dass Gillissen den oben genannten SGRQ als am häufigsten verwendeten Fragebogen zur Erfassung der Lebensqualität in COPD-Studien bezeichnet [20]. Diese Skala wird also auch ohne Änderung des Berichtsplans in die Nutzenbewertung einfließen.

Daher ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans.

2.3.6 Lungenfunktion

In mehreren Stellungnahmen wurde empfohlen, die Lungenfunktion als Endpunkt in die Nutzenbewertung aufzunehmen. Als Messgrößen wurden die Einsekundenkapazität (FEV_1) und die inspiratorische Kapazität (IC) genannt. Die FEV_1 sei der „am besten validierte Surrogatparameter für Mortalität und Exazerbationen“ [21,22]. Ein erniedrigter FEV_1-Wert sei „mit einer erhöhten Gesamtletalität assoziiert“ [20]. Die durch die IC zu erfassende Überblähung der Lunge sei „ein unabhängiger Prädikator für Mortalität bei COPD“ [23].

Die von den Stellungnehmenden genannte Literatur berichtet allerdings nur den prognostischen Wert dieser Messgrößen. Daraus lässt sich nicht ableiten, dass die Lungenfunktionsparameter bei der Bewertung von Interventionen valide Surrogate für patientenrelevante Endpunkte sind.

Die IQWiG-Methoden fordern eindeutige Belege aus Interventionsstudien, dass es einen plausiblen, starken, konsistenten und gleichgerichteten Zusammenhang zwischen der Änderung des Surrogats und der Änderung des patientenrelevanten Endpunkts gibt [7]. Anders ausgedrückt, bedeutet dies, dass der (Interventions-) Effekt auf das Surrogat den (Interventions-) Effekt auf den patientenrelevanten Endpunkt (in ausreichendem Maß) erklären muss. Ein Surrogat kann somit prinzipiell nicht als valide betrachtet werden, wenn es Studien gibt, in denen eine Intervention einen inkonsistenten Einfluss auf den Surrogat- und einen patientenrelevanten Endpunkt hat oder in denen noch nicht einmal eine starke Korrelation zwischen Surrogat- und patientenrelevanten Endpunkt gezeigt werden konnte. Beides trifft jedoch für die folgenden Beispiele zu.

In den Studien zu Interventionen bei COPD wurden wiederholt inkonsistente Effekte auf die Lungenfunktion und patientenrelevante Endpunkte gezeigt. So stellten z. B. die Autoren in einer Studie mit Oxitropium vs. Placebo fest, dass sich zwar Verbesserungen bei der Belastbarkeit (6MWT) und COPD-Symptomen (Borg-Skala) nach Oxitropiumgabe zeigen lassen, dass deren Ausmaß aber nicht mit der Zunahme der forcierten Vitalkapazität (FVC) und der FEV_1 nach Studienmedikation zusammenhängt [24]. In einer Studie mit Ipratropiumbromid vs. Placebo zeigte sich, dass sich nach Ipratropiumbromid ein Belastungstest (symptom-limited exercise endurance time) und die zeitgleich erhobenen COPD-Symptome (Borg-Skala) signifikant verbesserten, die Änderungen aber nicht mit der individuellen FEV_1-Zunahme korrelierten [25]. In dieser Studie zeigte sich allerdings eine Korrelation zwischen der IC und den Ergebnissen des Belastungstests.

Auch in Querschnittsstudien zeigten sich nur schwache Korrelationen zwischen der FEV_1 und patientenrelevanten Endpunkten. So fanden die Autoren in einer Regressionsanalyse ihrer Messungen bei 143 COPD-Patienten keinen Beitrag der FEV_1 zur Varianz von 3 verglichenen Lebensqualitätsfragebögen [26]. Die gleiche Autorengruppe setzte in einer weiteren Publikation die COPD-Schweregrade von 194 COPD-Patienten auf Basis der FEV_1 in Relation zu den entsprechenden Werten der gesundheitsbezogenen Lebensqualität der Patienten, wobei sich die Schweregrade II und III nicht mit dem SGRQ diskriminieren ließen [27].

Zusammenfassend kommt auch die ATS/ERS-Task-Force „Outcomes for COPD pharmacological trials“ zu dem Schluss, dass die FEV_1 nach Bronchodilatatorgabe nur schwach mit patientenrelevanten Endpunkten wie Dyspnoe, Belastbarkeit und gesundheitsbezogene Lebensqualität korreliert [9]. Insgesamt ist also die Lungenfunktion kein valides Surrogat für patientenrelevante Endpunkte und deshalb für die Nutzenbewertung nicht geeignet.

In der Diskussion dieses Punktes in der mündlichen Erörterung wurde erneut lediglich der prognostische Wert dieser Größen hervorgehoben. Darüber hinaus wurde eine Tiotropiumbromidstudie genannt, bei der, wie bei der oben bereits erwähnten Studie (ebenfalls publiziert von O'Donnell et al.), die Effekte von Tiotropiumbromid für die IC mit denen für Belastungstests korrelierten [28].

Grundsätzlich wird der Stellenwert der Lungenfunktion, insbesondere der FEV_1, als valides Surrogat für patientenrelevante Endpunkte in Studien zur COPD in der Literatur infrage gestellt. Dies ist durch inkonsistente Effekte aus Interventionsstudien und fehlende Korrelationen aus Querschnittsstudien für die Lungenfunktion und patientenrelevante Endpunkte begründet. Von den Stellungnehmenden wurde keine Literatur vorgelegt, die eine Validität der IC als Surrogat für patientenrelevante Endpunkte, z. B. im Sinne der von Weir et al. beschriebenen Methoden [29], belegt. Aus diesem Grund kann auch die IC trotz der Beobachtung gleichgerichteter Effekte bei der IC und Belastungstests in einzelnen Studien nicht als valides Surrogat gelten. Daher wird keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans gesehen.

2.4 Nutzenbewertung auf alleiniger Basis von randomisierten kontrollierten Studien (RCTs)

In einer Stellungnahme wurde kritisiert, dass die Nutzenbewertung allein auf Basis randomisierter kontrollierter Studien durchgeführt wird. Begründet wurde dies damit, dass das Prinzip der bestverfügbaren Evidenz anzuwenden sei. Mit GRADE [30] liege ein erprobtes und bewährtes Instrument für eine differenzierte Bewertung von Gesamtevidenz unterschiedlicher Evidenzstufen vor. Zudem sei der Hinweis, dass RCTs möglich und praktisch durchführbar sind, nicht hilfreich, denn entscheidend sei das Vorliegen solcher Studien. Es erscheine „dem Kosten-Nutzen-Verhältnis des Erkenntnisgewinns nicht angemessen, im Falle des Nicht-Vorliegens von RCTs, vorhandene relevante nicht randomisierte Studien komplett und a priori geplant zu ignorieren". Auch bei der Beurteilung von Schadensmerkmalen von Interventionen könne durch das Einbeziehen von nicht randomisierten Studien zusätzliche Evidenz zu einer Gesamtbeurteilung beitragen– entweder in Ergänzung zu RCTs oder – bei deren Nichtverfügbarkeit – als alleinige Evidenzbasis.

Nach der Definition von Sackett für evidenzbasierte Medizin (EbM) wird vom Arzt im Rahmen der Behandlung eines einzelnen Patienten die „beste verfügbare externe Evidenz" herangezogen und auf den spezifischen, individuellen Einzelfall einschließlich der Berücksichtigung der Patientenerfahrungen und -präferenzen angewendet. Sackett schreibt zwar im gleichen Artikel, dass EbM nicht auf randomisierte kontrollierte Studien und Meta-Analysen begrenzt sei, erläutert dies jedoch fragebezogen so: „Insbesondere bei der Frage nach Therapiemethoden sollten wir jedoch nicht-experimentelle Ansätze vermeiden, da diese häufig zu falsch-positiven Schlüssen hinsichtlich der Wirksamkeit von Maßnahmen kommen." [31] Die Aufgabe des IQWiG ist von der des Arztes (individuelle

Fallentscheidung) zu unterscheiden, denn die Ergebnisse der geplanten Nutzenbewertung können Einfluss auf die Versorgung von Patienten in Deutschland haben. Aus diesem Grund müssen hohe Anforderungen an die Ergebnissicherheit der Studien, die in die Nutzenbewertung einfließen, gestellt werden. Das Studiendesign hat aber erheblichen Einfluss auf die Ergebnissicherheit. So kann mit Beobachtungsstudien, prospektiv oder retrospektiv, ein kausaler Zusammenhang zwischen Intervention und Effekt in der Regel nicht dargestellt werden, während die kontrollierte Interventionsstudie grundsätzlich hierfür geeignet ist [32]. Dies gilt vor allem dann, wenn andere, das Ergebnis beeinflussende Faktoren ganz oder weitgehend ausgeschaltet werden. Aus diesem Grund stellt die randomisierte kontrollierte Studie den Goldstandard bei der Bewertung medikamentöser und nichtmedikamentöser Interventionen dar [33].

Im Fall von Tiotropiumbromid ist das Vorliegen von RCTs nicht nur möglich, sondern zudem durch das Zulassungsverfahren sichergestellt. Zur Beurteilung von Schadensmerkmalen ist anzumerken, dass unerwünschte Ereignisse und damit Aspekte der Sicherheit und Verträglichkeit gemäß Good Clinical Practice in jeder Interventionsstudie und damit auch in jeder RCT erhoben werden müssen [34].

Auch die Anwendung umfangreicher Bewertungssysteme löst nicht das Problem, dass der Einschluss von Studien niedrigerer Evidenz die Gefahr der Verzerrung des Nutzenbewertungsergebnisses birgt.

Daher lässt keines der vorgebrachten Argumente aus Sicht des Instituts die Einbeziehung nicht randomisierter Studien als notwendig erscheinen, sodass sich diesbezüglich kein Änderungsbedarf für den Berichtsplan ergibt.

Die Methoden des Instituts sehen vor, dass in der Anhörung zum vorläufigen Berichtsplan nicht randomisierte Studien eingereicht werden können, wenn der Berichtsplan die Informationsbeschaffung auf RCTs beschränkt. In solchen Fällen ist aber zusätzlich eine adäquate Begründung der Validität der kausalen Interpretation der in diesen Studien beschriebenen Effekte erforderlich. Von den Stellungnehmenden wurden keine nicht randomisierten Studien benannt oder eingereicht.

Davon unabhängig besteht im Rahmen des Stellungnahmeverfahrens zum Vorbericht ebenfalls die Möglichkeit, auf qualitativ angemessene Studien zu verweisen, die aus Sicht des jeweiligen Stellungnehmenden eine valide Beantwortung der Fragestellung des Berichts ermöglichen. Auch hier ist bei der Einreichung nicht randomisierter Studien eine adäquate Begründung der Validität der kausalen Interpretation der in diesen Studien beschriebenen Effekte notwendig.

2.5 Studiendauer

Im Berichtsplan wird eine minimale Studiendauer von 6 Monaten festgelegt. Diese Studiendauer wurde in mehreren Stellungnahmen diskutiert.

Ein Stellungnehmender betonte den besonderen Wert von Langzeitstudien von 2 und mehr Jahren für die Nutzenbewertung. Als Gründe wurden saisonale Unterschiede und die zu kleine Anzahl von Ereignissen bei Studien kürzerer Dauer genannt.

Andere Stellungnehmende merkten an, dass die European Medicines Agency (EMA) lediglich fordere, dass eine Studie mindestens 6 Monate lang sein soll, nicht aber, dass alle Studien diese Dauer haben sollen [1]. Es wurde außerdem darauf hingewiesen, dass auch in kürzeren Studien patientenrelevante Effekte auftreten könnten und dass solche Effekte nicht negiert werden sollten, solange nicht Evidenz für transiente Effekte vorliege. Eine Stellungnahme wies darauf hin, dass insbesondere für Lungenfunktionsparameter kürzere Studien deutliche Effekte zeigen könnten.

In der vorliegenden Nutzenbewertung sollen patientenrelevante Endpunkte betrachtet werden. Dabei handelt es sich im Wesentlichen um COPD-Symptomatik einschließlich körperlicher Belastbarkeit und Exazerbationen, gesundheitsbezogener Lebensqualität und COPD-bedingter bzw. -assoziierter Mortalität.

Die Formulierung der EMA lässt aus Sicht des IQWiG nicht den Schluss zu, dass beim Vorliegen einer 6-Monats-Studie zur Bewertung des Effekts bezüglich der COPD-Symptome weitere Studien von kürzerer Dauer für diesen Endpunkt berücksichtigt werden. Die Empfehlungen der Food and Drug Administration (FDA) sprechen für die Untersuchung von Effekten zur Linderung von Symptomen generell von einer Studiendauer von mindestens 6 Monaten [8]. Für Studien zur Untersuchung der Effekte auf Exazerbationen verlangt die FDA Studien von mindestens 1 Jahr.

Bezüglich der Bewertung der gesundheitsbezogenen Lebensqualität wurde in einer Stellungnahme ein weiteres Dokument der EMA zitiert [35], in dem bei der Bestimmung der Lebensqualität eine Studiendauer von 3 bis 6 Monaten empfohlen wird. Bei dieser Empfehlung handelt es sich allerdings um ein nicht indikationsspezifisches Papier, während sich andererseits das im Berichtsplan zitierte Dokument [1] explizit auf COPD bezieht und auch konkret die Lebensqualitätsskala St. George’s Respiratory Questionnaire nennt.

Da der Auftrag aus der Nutzenbewertung einer Dauertherapie besteht, müsste bei der Bewertung von Studien mit einer Dauer von z. B. nur 12 Wochen sichergestellt sein, dass dieser Effekt auch über eine längere Dauer konstant anhält. Zu dieser Frage wurde von den Stellungnehmenden keine Literatur vorgelegt.

In der Literatur gibt es vereinzelte Beispiele für transiente Effekte patientenrelevanter Endpunkte in Interventionsstudien mit COPD-Patienten. So zeigte [36] eine Post-hoc-Subgruppenanalyse eines Pools zweier Studien einen transienten Effekt von Salmeterol für die Dyspnoe [36]. Während in Woche 8 der Behandlung ein statistisch signifikanter Effekt von Salmeterol vs. Placebo beobachtet wurde, war dieser in Woche 16 und 24 nicht mehr sichtbar. [37]In einer Post-hoc-Subgruppenanalyse eines Pools zweier anderer Studien wurden ebenfalls unterschiedliche Effekte für die Dyspnoe im Lauf der Studie beobachtet. Tiotropiumbromid zeigte einen statistisch signifikanten Effekt im Vergleich zu Placebo in Woche 6, 25 und 49, nicht jedoch in Woche 12 und 36 [37].

In der Erörterung wurde von den Stellungnehmenden darauf hingewiesen, dass insbesondere in Studien mit Patienten mit schwerer COPD auch bei kürzerer Studiendauer ausreichende Ereignisraten von Exazerbationen erreicht werden können, um Effekte von Bronchodilatatoren zu messen. Dieser Hinweis beantwortet nicht die Frage, ob diese Effekte über eine längere Zeit andauern würden.

Zusammenfassend erscheint eine Beschränkung auf Studien von mindestens 6 Monaten Dauer für die vorliegende Nutzenbewertung angemessen. Daher ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans.

2.6 Bewertung der Relevanz von Effekten

In Abschnitt 4.4 des vorläufigen Berichtsplans wurde beschrieben, dass bei Betrachtung patientenrelevanter Endpunkte, die mithilfe von (komplexen) Skalen operationalisiert werden, die Notwendigkeit besteht, neben der statistischen Signifikanz die Relevanz der Effekte zu bewerten. Es wurde ein abgestuftes Verfahren dargestellt, wie diese Relevanzbewertung in Abhängigkeit von der Verfügbarkeit verschiedener Relevanzkriterien konkret im Rahmen der Nutzenbewertung durchgeführt werden soll.

In mehreren Stellungnahmen wurden insbesondere 3 Punkte dieses Vorgehens kritisiert. Zum einen hinterfragten die Stellungnehmenden die Notwendigkeit, überhaupt eine Bewertung der Relevanz von Effekten vorzunehmen. Zum anderen wurde die Verwendung standardisierter Mittelwertdifferenzen (SMD) zur Relevanzbewertung als ungeeignet eingestuft. Schließlich sei die Anwendung statistischer Tests mit verschobener Nullhypothese (Tests auf relevante Überlegenheit) zur Relevanzbewertung aus methodischen Gründen zweifelhaft.

Notwendigkeit der Bewertung der Relevanz

In den meisten Stellungnahmen wurde argumentiert, dass bei patientenrelevanten Endpunkten jeder statistisch signifikante Unterschied als relevant einzustufen sei. In einer Stellungnahme wurde in ähnlicher Weise gefordert, dass zumindest bei patientenrelevanten Endpunkten, für die keine ankerbasierte Relevanzschwelle existiert, die statistische Signifikanz ausreichend für einen Nutzenbeleg sei. Des Weiteren forderten einige Stellungnehmende, zur Frage nach

Relevanzkriterien sei klinischer Sachverstand einzubinden. Diese Forderung zeigt, dass die Notwendigkeit der Bewertung der Relevanz von Effekten z. T. auch vonseiten der Stellungnehmenden gesehen wird. Diese Notwendigkeit findet sich auch in verschiedenen Richtlinien von Zulassungsbehörden. Zum Beispiel wird bezogen auf Lebensqualitätsskalen [35] und auf patientenberichtete Endpunkte [38] von der europäischen sowie der US-amerikanischen Zulassungsbehörde festgestellt, dass statistisch signifikante Effekte u. U. nicht relevant seien und somit für solche Endpunkte die Relevanz der Ergebnisse zu bewerten sei. Insbesondere im Rahmen der Nutzenbewertungen des Instituts besteht durch die erhöhte Trennschärfe (Power) aufgrund der meta-analytischen Zusammenfassung der Ergebnisse mehrerer Studien die Gefahr, statistisch signifikante, jedoch irrelevante Effekte aufzudecken.

Für das Institut besteht daher weiterhin die Notwendigkeit, bei Ergebnissen, die auf (komplexen) Skalen beruhen, die Relevanz zu bewerten.

Die Forderung, klinischen Sachverstand zur Frage nach Relevanzkriterien einzubeziehen, unterstützt das Institut. Diese Einbeziehung ist zum einen durch die Verfahrensabläufe des Instituts bei Nutzenbewertungen sichergestellt. Zum anderen beschreibt der Berichtsplan explizit, dass zur Relevanzbewertung primär auf validierte Relevanzkriterien (Relevanzschwellen, Minimal Important Differences [MID]), die ohne klinischen Sachverstand gar nicht zu ermitteln sind, zurückgegriffen werden wird. Nur dann, wenn solche instrumentspezifischen Relevanzschwellen nicht vorliegen, bewertet das IQWiG die Relevanz anhand allgemeiner Relevanzschwellen.

Verwendung standardisierter Mittelwertdifferenzen (SMDs) zur Relevanzbewertung

In den Stellungnahmen wurde die Wahl von SMDs als Effektmaß zur Relevanzbewertung als ungeeignet eingestuft. Dieses Effektmaß sei rein verteilungsbasiert und könne aufgrund des fehlenden klinischen Ankers nicht herangezogen werden. In einer Stellungnahme wurde darüber hinaus darauf hingewiesen, dass die im Berichtsplan festgelegte Relevanzschwelle von 0,2 für dieses Effektmaß als fixe Grenze nicht adäquat sei, da je nach Kontext auch kleinere Werte relevant sein könnten.

Das Institut teilt die Ansicht, dass eine auf der SMD basierende Relevanzbewertung keinen direkten klinischen Anker berücksichtigt. Es hält die eventuelle Verwendung der SMD im Rahmen des beschriebenen abgestuften Vorgehens aus mehreren Gründen dennoch für richtig.

Im Rahmen von statistischen Auswertungen klinischer Studien finden sich grundsätzlich Festlegungen ohne Berücksichtigung des konkreten klinischen Kontexts. Das Signifikanzniveau wird beispielsweise fast immer ohne Beachtung spezieller klinischer Situationen auf 5 % festgesetzt. Des Weiteren sei angemerkt, dass beim Vergleich zweier Gruppen bezüglich eines stetigen Endpunkts die statistische Signifikanz üblicherweise anhand von SMDs geprüft wird (siehe die Teststatistik des t-Tests). D. h. es ist durchaus üblich,

klinische Entscheidungen rein statistisch und ohne Beachtung des Kontexts zu treffen. Dennoch teilt das Institut die Ansicht, nach Möglichkeit klinische Sachverhalte in seine Bewertungen einfließen zu lassen. Das spiegelt sich auch in dem beschriebenen abgestuften bzw. hierarchischen Vorgehen wider. Sollten für einen Endpunkt validierte oder zumindest etablierte Relevanzkriterien vorliegen, werden diese für die Relevanzbewertung verwendet. Das können skalen- und indikationsspezifische Relevanzschwellen für Gruppenunterschiede oder auf MIDs beruhende Responderanalysen sein. Nur dann, wenn für keines dieser Relevanzkriterien solche Informationen vorliegen sollten, wird das Institut die Relevanz anhand der SMD bewerten. Da das Institut eine wie auch immer geartete Relevanzbewertung bei (komplexen) Skalen als notwendig erachtet (siehe oben), bestünde anderenfalls (beim Fehlen skalenspezifischer Relevanzkriterien) keine Möglichkeit, statistisch signifikante Effekte auf ihre Relevanz hin zu überprüfen. Die Betrachtung der SMD liefert somit eine weitere Möglichkeit, sich einer Relevanzbewertung zu nähern.

Die Kritik, dass eine feste und kontextunabhängige Relevanzgrenze von 0,2 für die SMD nicht jeder klinischen Situation gerecht werden kann, ist nachvollziehbar. Aber auch hier sei darauf verwiesen, dass die Anwendung dieser Grenze die letzte Option im Rahmen des abgestuften Vorgehens darstellt. Sollten skalen- und indikationsspezifische Relevanzkriterien vorliegen, findet die Grenze von 0,2 SMD keine Anwendung.

Anwendung statistischer Tests mit verschobener Nullhypothese

In einigen Stellungnahmen wurde die Anwendung statistischer Tests mit verschobener Nullhypothese (in Richtung der zu prüfenden Intervention) bzw. die dazu äquivalente Forderung, dass das zum beobachteten Effekt korrespondierende Konfidenzintervall vollständig oberhalb der Relevanzschwelle liegt, kritisiert. Dies wurde in einigen Stellungnahmen mit den dafür notwendigen sehr großen Fallzahlen begründet. In Folge könne eine durchaus relevante MID praktisch nicht nachgewiesen werden. Andere Stellungnehmende argumentierten in ähnlicher Weise, dass die vom Institut gewählte Relevanzschwelle von 0,2 SMD einem „wünschenswerten und (…) nicht zu übersehenden Effekt (δ_{rel})“ entspreche und nicht einem niedriger zu wählenden „kleinsten Effekt (δ_{min})“.

Hier liegt offenbar ein Missverständnis in Bezig auf die Begriffe vor. Grundsätzlich muss unterschieden werden zwischen MIDs für individuelle Veränderungen eines Patienten, die nur der Bildung von Responderdefinitionen dienen können, und Relevanzschwellen für Gruppenunterschiede. Erstere sind nicht mit Letzteren gleichzusetzen [39]. Des Weiteren stellt die im Berichtsplan genannte Relevanzschwelle, die der verschobenen Hypothesengrenze entspricht, den Grenzwert zwischen einem irrelevanten und einem relevanten Gruppenunterschied dar und ist somit nicht als „durchaus relevant“ zu deuten. Bezogen auf die oben zitierte Stellungnahme ist die vom Institut als Relevanzschwelle bezeichnete Grenze mit δ_{min} gleichzusetzen und nicht mit δ_{rel}.

Insgesamt sieht das Institut keine Notwendigkeit, vom im vorläufigen Berichtsplan beschriebenen hierarchischen Vorgehen abzuweichen. Um mögliche Missverständnisse zu vermeiden, wurde die Beschreibung dieses Vorgehens jedoch sprachlich überarbeitet.

2.7 Ausschluss von Ergebnissen

2.7.1 Ausschluss von Ergebnissen auf Basis von weniger als 70 % der auszuwertenden Patientendaten

Im vorläufigen Berichtsplan wird ausgeführt, dass in bestimmten Fällen einzelne Ergebnisse aus den Studien zu einem Endpunkt nicht einbezogen werden, insbesondere wenn viele Patienten nicht in der Auswertung enthalten sind. Dies gilt in der Regel dann, wenn die Ergebnisse auf weniger als 70 % der in die Auswertung einzuschließenden Patienten basieren, d. h. der Anteil der fehlenden Werte größer als 30 % ist.

In einer Stellungnahme wurde angemerkt, dass es nicht gerechtfertigt sei, dass eine Abbruchrate von mehr als 30 % zum Ausschluss der jeweiligen Ergebnisse führe. Hierbei wurde auch das Ersetzungsverfahren Last Observation Carried Forward (LOCF) als adäquate Vorgehensweise angeführt, d. h. das Fortschreiben des letzten verfügbaren Werts zu einem Endpunkt unter Behandlung für etwaige Studienabbrecher. Ein anderer Stellungnehmender regte an, klarzustellen, dass sich die genannten Zahlen nicht auf die vorzeitigen Abbruchraten beziehen, sondern auf den Anteil der Patienten, die in eine Analyse eingeschlossen werden. Zugleich wurde darum gebeten, anzugeben, welches „die Basis (der Nenner) bei der Berechnung dieser Kriterien“ darstellt.

Wie oben bereits ausgeführt, lautet die Formulierung im Berichtsplan „auf weniger als 70 % der *in die Auswertung einzuschließenden Patienten* basieren“. Das Verfahren bezieht sich also nicht auf Abbruchraten, denen mit geeigneten Ersetzungsverfahren begegnet wird. Vielmehr betrifft das Kriterium Situationen, in denen mehr als 30 % der Patienten keinerlei Berücksichtigung in der Analyse eines Endpunkts finden. Dies wurde auch in der Erörterung auf Nachfrage eines Teilnehmers noch einmal erläutert.

Die Basis (der Nenner) bei der Berechnung dieser Kriterien bildet i. d. R. die Anzahl der in die jeweiligen Gruppen randomisierten Patienten. In begründeten Einzelfällen kann sich die Zahl der auszuwertenden Patienten verringern, wenn klar ersichtlich ist, dass das Kriterium, nach dem Patienten aus der Analyse ausgeschlossen werden, unabhängig vom Therapieverlauf und von der Gruppenzugehörigkeit ist. Im Berichtsplan wird in Abschnitt 4.4.1 eine solche Konstellation verdeutlicht durch das Beispiel, dass aus logistischen Gründen für ganze Zentren (ganze Randomisierungsblöcke) keine Daten erhoben wurden und dies bereits bei der Studienplanung vorgesehen war.

Zusammenfassend ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des im Berichtsplan beschriebenen Verfahrens. Um mögliche Missverständnisse zu vermeiden, wurde die Beschreibung des Vorgehens jedoch sprachlich überarbeitet.

2.7.2 Ausschluss von Ergebnissen: weitere Gründe

Eine Stellungnahme merkte an, dass in Abschnitt 4.4.1 ausgeführt werde, dass für die Nutzenbewertung einzelne patientenrelevante Endpunkte nicht berücksichtigt würden, „insbesondere“ wenn „viele Patienten nicht in der Auswertung enthalten sind“, und bat darum, im Berichtsplan klar zu definieren, welche weiteren Gründe zur Nicht-berücksichtigung von Endpunkten führen können.

Endpunkte werden z. B. auch dann nicht berücksichtigt, wenn sie mit nicht validierten Skalen erhoben wurden. Es ist aber sinnvoll, im Berichtsplan nur Methodenkriterien für vorhersehbare Konstellationen zu nennen; zudem ist eine erschöpfende Aufzählung aller Eventualitäten, die eine Spezifizierung der Methodik erfordern, auch nicht möglich. Sollten aber solche Konstellationen auftreten, wird die angepasste Methodik im Vorbericht dargelegt und steht daraufhin in der Anhörung zum Vorbericht zur Diskussion, sodass eine Stellungnahmemöglichkeit auch in einem solchen Fall gewährleistet ist.

Daher ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans.

2.8 Verwendung standardisierter Mittelwertdifferenzen

Zur Frage der Verwendung standardisierter Mittelwertdifferenzen (SMDs) wurden, abgesehen vom Kontext der Relevanzbewertung, in einer Stellungnahme noch einige andere Aspekte hinterfragt, die im Folgenden diskutiert werden.

Es wurde für notwendig gehalten, dass im Berichtsplan festgelegt wird, in welchen Situationen Effektschätzer aus Einzelstudien und deren Standardabweichungen in einer Meta-Analyse mittels SMDs gemeinsam ausgewertet werden. Es wurde zudem empfohlen, eine SMD-Betrachtung nur zusätzlich zur Betrachtung der Originalskala und nicht alternativ vorzunehmen. Darüber hinaus solle geklärt werden, „welche Variante von Hedges’ g bei dieser Nutzenbewertung eingesetzt werden soll, da verschiedene Varianten der Adjustierung von Cohen’s d existieren“.

Die Wahl des Distanzmaßes im Rahmen der Meta-Analysen hängt maßgeblich von der Art und Anzahl der bezüglich einer Fragestellung zusammenfassbaren Ergebnisse ab. Basieren die Ergebnisse für einen Endpunkt fast ausnahmslos auf derselben Skala, dann wird i. d. R. die unstandardisierte Differenz der Mittelwerte (im Falle stetiger Ergebnisse) zur Prüfung der statistischen Signifikanz herangezogen. Werden zwischen den zusammenzufassenden Studien verschiedene Skalen für einen Endpunkt verwendet, bleibt für die Meta-Analysen nur die

Wahl der standardisierten Differenzen. In jedem Fall werden, ggf. zusätzlich, SMDs bei der Bewertung der Relevanz von Skalen verwendet.

Zur Berechnung der SMDs in Form von Hedges' g wird die Formel, die für Cochrane-Reviews Verwendung findet, herangezogen [40].

Das vom IQWiG gewählte Vorgehen entspricht damit dem Standardvorgehen in den beschriebenen Situationen. Aus Sicht des IQWiG besteht deshalb keine Notwendigkeit einer detaillierteren Beschreibung im Berichtsplan.

Zusammenfassend ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans.

2.9 Meta-Analysen beim Vorliegen verschiedener statistischer Auswertetechniken

Im Berichtsplan sollte laut einer Stellungnahme festgelegt werden, wie in Meta-Analysen mit dem Vorliegen unterschiedlicher Auswertetechniken in Einzelstudien umgegangen werden soll. Genannt wurde u. a. das Vorhandensein oder Fehlen einer Adjustierung nach Kovariablen.

Dieser Aspekt wird in Abschnitt 4.4.2 des Berichtsplans behandelt, in dem es heißt, dass für die statistische Auswertung primär die Ergebnisse aus Intention-to-Treat-Analysen, *so wie sie in den vorliegenden Dokumenten beschrieben sind*, verwendet werden. Eine umfassende Darstellung der möglichen Spezifizierungen, die alle Eventualitäten einschließt, ist nicht möglich.

Daher ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans.

2.10 Berücksichtigung der Inhalationssysteme

2.10.1 Bewertung verschiedener Applikationsformen von Tiotropiumbromid

In einer Stellungnahme wurde angeregt, im Berichtsplan anzugeben, wie beim Teilziel des Vergleichs von Tiotropiumbromid mit Therapiealternativen verfahren wird, falls sich eine Differenzierung der beiden verfügbaren Tiotropiumbromid-Inhalatoren bezüglich des Nutzens oder Schadens ergibt.

Es wurde vorgeschlagen, dass in einem solchen Fall die zu prüfende Intervention getrennt (stratifiziert) nach den beiden verfügbaren Inhalatoren für Tiotropiumbromid mit den Vergleichstherapien verglichen und bewertet werden solle.

Der Einfluss der beiden Tiotropiumbromid-Anwendungsformen HandiHaler und Respimat kann in direkt vergleichenden Studien untersucht werden. Darüber hinaus ist es möglich, zu prüfen, ob der Inhalatortyp in einem Pool aus Studien mit unterschiedlichen Inhalatoren zu

Heterogenität führt und diese erklärt. Der Berichtsplan beschreibt die in dieser Situation anzuwendenden Methoden in Abschnitt 4.4.4.

Der Inhalatortyp wurde aber an dieser Stelle nicht noch mal als zu untersuchender potenzieller Effektmodifikator aufgeführt, da sich dies implizit aus der Zielformulierung in Kapitel 2 ergibt. Zudem wird dieser Fall auch durch die folgende Formulierung in Abschnitt 4.4.4 abgedeckt: „Sollten sich aus den verfügbaren Informationen Anhaltspunkte für weitere mögliche Effektmodifikatoren ergeben, können diese ebenfalls begründet einbezogen werden."

Daher ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans.

2.10.2 Bewertung verschiedener Applikationsformen der Vergleichsinterventionen

In einer Stellungnahme wurde vorgeschlagen, im Berichtsplan klarer darzustellen, wie beim Vergleich von Tiotropiumbromid mit Therapiealternativen mit verschiedenen Inhalatorsystemen in der Vergleichstherapie verfahren werden soll. Die Stellungnehmenden halten den Faktor Inhalatorsystem für die Bewertung für relevant und regten an, die Frage der Inhalatorsysteme im Sinne von Subgruppenanalysen einfließen zu lassen.

Der Berichtsplan beschreibt in Abschnitt 4.4.4 das Vorgehen bei der Analyse von Subgruppenmerkmalen und anderen Effektmodifikatoren. Spezifika der Behandlung (und darum handelt es sich bei einem Inhalatorsystem) werden explizit als mögliche Effektmodifikatoren genannt. Der Abschnitt stellt klar, dass und wie mögliche Effektunterschiede untersucht werden sollen. Es wird darüber hinaus beschrieben, dass immer dann, wenn sich aus den verfügbaren Informationen Anhaltspunkte dafür ergeben, dass über die explizit genannten Subgruppenmerkmale hinausgehende Effektmodifikatoren vorliegen, diese einbezogen werden.

Sollte es sich im Laufe der Bewertung herausstellen, dass der Inhalatortyp einer Vergleichsintervention ein Effektmodifikator sein könnte, so wird dies mit den in Abschnitt 4.4.4 des Berichtsplans beschriebenen Verfahren untersucht.

Daher ergibt sich keine Notwendigkeit zur Änderung oder Ergänzung des Berichtsplans.

2.11 Literaturverzeichnis

1. European Medicines Agency. Points to consider on clinical investigation of medicinal products in the chronic treatment of patients with chronic obstructive pulmonary disease (COPD) [online]. 19.05.1999 [Zugriff: 29.10.2009]. URL: http://www.emea.europa.eu/pdfs/human/ewp/056298en.pdf.

2. Abholz HH, Gillissen A, Magnussen H, Schott G, Schultz K, Ukena D et al. Nationale VersorgungsLeitlinie COPD: Langfassung; Version 1.6 [online]. 04.2008 [Zugriff: 04.06.2009]. URL: http://www.versorgungsleitlinien.de/themen/copd/pdf/nvl_copd_lang.pdf.

3. American Thoracic Society, European Respiratory Society. Standards for the diagnosis and management of patients with COPD [online]. 2004 [Zugriff: 13.07.2010]. URL: http://www.thoracic.org/clinical/copd-guidelines/resources/copddoc.pdf.

4. Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of Chronic Obstructive Pulmonary Disease [online]. 2009 [Zugriff: 07.12.2009]. URL: http://www.goldcopd.com/download.asp?intId=548.

5. Vogelmeier C, Buhl R, Criée CP, Gillissen A, Kardos P, Köhler D et al. Leitlinie der Deutschen Atemwegsliga und der Deutschen Gesellschaft für Pneumologie und Beatmungsmedizin zur Diagnostik und Therapie von Patienten mit chronisch obstruktiver Bronchitis und Lungenemphysem (COPD). Pneumologie 2007; 61(8): e1-e40.

6. Casaburi R, Kukafka D, Cooper CB, Witek TJ Jr, Kesten S. Improvement in exercise tolerance with the combination of tiotropium and pulmonary rehabilitation in patients with COPD. Chest 2005; 127(3): 809-817.

7. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden: Version 3.0 [online]. 27.05.2008 [Zugriff: 08.10.2008]. URL: http://www.iqwig.de/download/IQWiG_Methoden_Version_3_0.pdf.

8. Food and Drug Administration. Guidance for industry: chronic obstructive pulmonary disease; developing drugs for treatment; draft guidance [online]. 11.2007 [Zugriff: 30.10.2009]. URL: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071575.pdf.

9. Cazzola M, MacNee W, Martinez FJ, Rabe KF, Franciosi LG, Barnes PJ et al. Outcomes for COPD pharmacological trials: from lung function to biomarkers. Eur Respir J 2008; 31(2): 416-469.

10. Nathan RA, Sorkness CA, Kosinski M, Schatz M, Li JT, Marcus P et al. Development of the asthma control test: a survey for assessing asthma control. J Allergy Clin Immunol 2004; 113(1): 59-65.

11. GlaxoSmithKline. COPD assessment test [online]. 2009 [Zugriff: 06.07.2010]. URL: http://www.catestonline.org/images/pdfs/GermanCATest.pdf.

12. Salzman SH. The 6-min walk test: clinical and research role, technique, coding, and reimbursement. Chest 2009; 135(5): 1345-1352.

13. American Thoracic Society. ATS statement: guidelines for the six-minute walk test. Am J Respir Crit Care Med 2002; 166(1): 111-117.

14. American Thoracic Society, American College of Chest Physicians. ATS/ACCP statement on cardiopulmonary exercise testing. Am J Respir Crit Care Med 2003; 167(2): 211-277.

15. Puente-Maestu L, Villar F, De Miguel J, Stringer WW, Sanz P, Sanz ML et al. Clinical relevance of constant power exercise duration changes in COPD. Eur Respir J 2009; 34(2): 340-345.

16. Jones PW. Health status and the spiral of decline. COPD 2009; 6(1): 59-63.

17. Jones PW, Quirk FH, Baveystock CM. The St George's Respiratory Questionnaire. Respir Med 1991; 85(Suppl B): 25-31.

18. Chassany O, Sagnier P, Marquis P, Fullerton S, Aaronson N. Patient-reported outcomes: the example of health-related qualtiy of life; a European guidance document for the improved integration of health-realted quality of life assessment in the drug regulatory process. Drug Inf J 2002; 36(1): 209-238.

19. Fayers P, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. Chichester: Wiley; 2007.

20. Gillissen A, Buhl R, Kardos P, Puhan M, Rabe KF, Rothe T et al. Studienendpunkte bei der chronisch-obstruktiven Lungenerkrankung (COPD): "minimal clinically important difference". Pneumologie 2008; 62(3): 149-155.

21. Celli BR, Cote CG, Marin JM, Casanova C, Montes de Oca M, Mendez RA et al. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. N Engl J Med 2004; 350(10): 1005-1012.

22. Donaldson GC, Wedzicha JA. COPD Exacerbations; 1: Epidemiology. Thorax 2006; 61(2): 164-168.

23. Casanova C, Cote C, De Torres JP, Aguirre-Jaime A, Marin JM, Pinto-Plata V et al. Inspiratory-to-total lung capacity ratio predicts mortality in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2005; 171(6): 591-597.

24. Hay JG, Stone P, Carter J, Church S, Eyre-Brook A, Pearson MG et al. Bronchodilator reversibility, exercise performance and breathlessness in stable chronic obstructive pulmonary disease. Eur Respir J 1992; 5(6): 659-664.

25. O'Donnell DE, Lam M, Webb KA. Spirometric correlates of improvement in exercise performance after anticholinergic therapy in chronic obstructive pulmonary disease. Am J Respir Crit Care Med 1999; 160(2): 542-549.

26. Hajiro T, Nishimura K, Tsukino M, Ikeda A, Koyama H, Izumi T. Comparison of discriminative properties among disease-specific questionnaires for measuring health-related quality of life in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 1998; 157(3 Pt 1): 785-790.

27. Hajiro T, Nishimura K, Tsukino M, Ikeda A, Oga T, Izumi T. A comparison of the level of dyspnea vs disease severity in indicating the health-related quality of life of patients with COPD. Chest 1999; 116(6): 1632-1637.

28. O'Donnell DE, Fluge T, Gerken F, Hamilton A, Webb K, Aguilaniu B et al. Effects of tiotropium on lung hyperinflation, dyspnoea and exercise tolerance in COPD. Eur Respir J 2004; 23(6): 832-840.

29. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. Stat Med 2006; 25(2): 183-203.

30. Guyatt G, Cook D, Jaeschke R, Schünemann H, Pauker S. Moving from evidence to action: grading recommendations; a qualitative approach. In: Guyatt G, Rennie D (Ed). Users´guides to the medical literature: a manual for evidence based practice. Chicago: AMA Press; 2002. S. 599-608.

31. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Was ist evidenz-basierte Medizin und was nicht? Munch Med Wochenschr 1997; 139(44): 644-645.

32. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet 2002; 359(9300): 57-61.

33. Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev 2007; (2): MR000012.

34. European Medicines Agency. ICH Topic E 6 (R1): guideline for good clinical practice; step 5 [online]. 07.2002 [Zugriff: 28.07.2010]. URL: http://www.emea.europa.eu/pdfs/human/ich/013595en.pdf.

35. European Medicines Agency. Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products [online]. 27.07.2005 [Zugriff: 06.07.2010]. URL: http://www.ema.europa.eu/pdfs/human/ewp/13939104en.pdf.

36. Hodder R, Kesten S, Menjoge S, Viel K. Outcomes in COPD patients receiving tiotropium or salmeterol plus treatment with inhaled corticosteroids. Int J Chron Obstruct Pulmon Dis 2007; 2(2): 157-167.

37. Adams SG, Anzueto A, Briggs DD Jr, Menjoge SS, Kesten S. Tiotropium in COPD patients not previously receiving maintenance respiratory medications. Respir Med 2006; 100(9): 1495-1503.

38. Food and Drug Administration. Guidance for industry: patient-reported outcome measures; use in medical product development to support labeling claims [online]. 12.2009 [Zugriff: 06.07.2010]. URL: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf.

39. Victor N. On clinically relevant differences and shifted null hypotheses. Methods Inf Med 1987; 26(3): 109-116.

40. Deeks JJ, Higgins JPT. Statistical algorithms in Review Manager 5 [online]. 11.2007 [Zugriff: 15.06.2010]. URL: http://www.cochrane.org/sites/default/files/uploads/Statistical_Methods_in_RevMan5.pdf.

3 Offenlegung potenzieller Interessenkonflikte

Im Folgenden sind die potenziellen Interessenkonflikte der Stellungnehmenden sowie weiterer Teilnehmer an der wissenschaftlichen Erörterung zusammenfassend dargestellt. Alle Informationen beruhen auf Selbstangabe der einzelnen Personen anhand des „Formblatts zur Offenlegung potenzieller Interessenkonflikte“. Das Formblatt ist unter www.iqwig.de abrufbar. Die in diesem Formblatt aufgeführten Fragen finden sich im Anschluss an diese Zusammenfassung.

3.1 Potenzielle Interessenkonflikte von Stellungnehmenden aus Organisationen, Institutionen und Firmen

Organisation / Institution / Firma	Name	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6
Boehringer Ingelheim Pharma GmbH & Co. KG	Glaab, Thomas, PD Dr.	ja	nein	nein	nein	nein	nein
	Kögler, Harald, PD Dr.	ja	nein	nein	nein	nein	nein
	Leimer, Inge, Dr.	ja	nein	nein	nein	nein	nein
Deutsche Atemwegsliga e. V.	Kardos, Peter, Dr.	nein	ja	ja	nein	nein	nein
GlaxoSmithKline GmbH & Co. KG	Banik, Norbert, Dr.	ja	ja	nein	nein	nein	ja
	Hagedorn, Cordula, Dr.	ja	ja	nein	nein	ja	nein
	Lützelberger, Uwe	ja	ja	nein	nein	nein	ja
Novartis Pharma GmbH	Färber, Lothar	ja	ja	nein	nein	ja	ja
	Krippner, Felix	ja	nein	nein	nein	nein	nein
	Peters, Lars	ja	nein	nein	nein	nein	nein
	Wasmuth, Timo	ja	ja	nein	nein	ja	ja
Pfizer Pharma GmbH	Hohmann, Christoph, Dr.	ja	nein	nein	nein	nein	ja
	Leverkus, Friedhelm	ja	nein	nein	nein	nein	ja
	Wernitz, Martin, Dr.	ja	nein	nein	nein	nein	nein

3.2 Potenzielle Interessenkonflikte von weiteren Teilnehmern an der wissenschaftlichen Erörterung (externe Sachverständige)

Name	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6
Bauer, Torsten, Prof. Dr.	nein	ja	ja	nein	nein	nein
Velasco Garrido, Marcial	nein	nein	nein	ja	ja	nein
Zentner, Annette, Dr.	nein	nein	nein	ja	ja	nein

Im „Formblatt zur Offenlegung potenzieller Interessenkonflikte“ wurden folgende 6 Fragen gestellt:

Frage 1: Sind oder waren Sie innerhalb des laufenden Jahres und der 3 Kalenderjahre davor bei einem Interessenverband im Gesundheitswesen oder einem vergleichbaren Interessenvertreter[6] abhängig (angestellt) beschäftigt? Falls ja, wo und in welcher Position?

Frage 2: Beraten Sie oder haben Sie innerhalb des laufenden Jahres und der 3 Kalenderjahre davor einen Interessenverband im Gesundheitswesen oder einen vergleichbaren Interessenvertreter direkt oder indirekt beraten? Falls ja, wen und wie hoch ist / war die Zuwendung / das Honorar?

Frage 3: Haben Sie abseits einer Anstellung oder Beratungstätigkeit innerhalb des laufenden Jahres oder der 3 Kalenderjahre davor im Auftrag eines Interessenverbands im Gesundheitswesen oder eines vergleichbaren Interessenvertreters Honorare für Vorträge, Stellungnahmen, Ausrichtung und / oder Teilnahme an Kongressen und Seminaren – auch im Rahmen von Fortbildungsveranstaltungen, oder für (populär-)wissenschaftliche oder sonstige Aussagen oder Artikel erhalten? Falls ja, von wem, für welche Tätigkeiten und wie hoch war die Zuwendung / das Honorar?

Frage 4: Haben Sie abseits einer Anstellung oder Beratungstätigkeit und / oder hat die Institution[7], bei der Sie angestellt sind bzw. die Sie vertreten, innerhalb des laufenden Jahres und der 3 Kalenderjahre davor von einem Interessenverband im Gesundheitswesen oder einem vergleichbaren Interessenvertreter finanzielle Unterstützung für Forschungsaktivitäten, andere wissenschaftliche Leistungen oder Patentanmeldungen erhalten? Falls ja, von wem, für welche Tätigkeit und in welcher Höhe?

Frage 5: Haben Sie und / oder hat die Institution, bei der Sie angestellt sind bzw. die Sie vertreten, innerhalb des laufenden Jahres oder der 3 Kalenderjahre davor sonstige finanzielle oder geldwerte Zuwendungen (z. B. Ausrüstung, Personal, Reisekostenunterstützung ohne wissenschaftliche Gegenleistungen) von einem Interessenverband im Gesundheitswesen oder einem vergleichbaren Interessenvertreter erhalten? Falls ja, von wem, aus welchem Anlass und in welcher Höhe?

Frage 6: Besitzen Sie Aktien, Optionsscheine oder sonstige Geschäftsanteile (auch in Fonds) von einer Firma oder Institution, die zu einem Interessenverband im Gesundheitswesen oder einem vergleichbaren Interessenvertreter gehört? Falls ja, von wem und welchen Wert haben diese aktuell?

---

[6] Dieses Formblatt erfasst finanzielle Beziehungen zu Interessenverbänden im Gesundheitswesen oder vergleichbaren Interessenvertretern, insbesondere der pharmazeutischen Industrie und der Medizinprodukteindustrie.

[7] Sofern Sie in einer ausgedehnten Institution tätig sind, ist es ausreichend, die geforderten Angaben auf Ihre Arbeitseinheit (z. B.: Klinikabteilung, Forschungsgruppe etc.) zu beziehen.

4 Dokumentation der wissenschaftlichen Erörterung – Teilnehmerliste, Tagesordnung und Protokoll

4.1 Teilnehmerliste der wissenschaftlichen Erörterung

Name	**Organisation / Institution / Firma / privat**
Banik, Norbert, Dr. Dr.	GlaxoSmithKline GmbH & Co. KG
Bauer, Torsten, Prof. Dr.	HELIOS Klinikum Emil von Behring
Fleer, Daniel, Dr.	IQWiG
Glaab, Thomas, PD Dr.	Boehringer Ingelheim Pharma GmbH & Co. KG
Hagedorn, Cordula, Dr.	GlaxoSmithKline GmbH & Co. KG
Hohmann, Christoph, Dr.	Pfizer Pharma GmbH
Kardos, Peter, Dr.	Deutsche Atemwegsliga e. V.
Kögler, Harald, PD Dr.	Boehringer Ingelheim Pharma GmbH & Co. KG
Lange, Stefan, PD Dr.	IQWiG (Moderation)
Leimer, Inge, Dr.	Boehringer Ingelheim Pharma GmbH & Co. KG
Leverkus, Friedhelm	Pfizer Pharma GmbH
Müller, Hildegard, Dr.	Sitzungsdokumentarischer Dienst, Landtag NRW (Protokollantin)
Peters, Lars	Novartis Pharma GmbH
Ringsdorf, Susanne	IQWiG
Skipka, Guido, Dr.	IQWiG
Velasco Garrido, Marcial	Technische Universität Berlin
Wernitz, Martin, Dr.	Pfizer Pharma GmbH
Wieseler, Beate, Dr.	IQWiG
Zentner, Annette, Dr.	Technische Universität Berlin

4.2 Tagesordnung der wissenschaftlichen Erörterung

	Begrüßung und Einleitung
TOP 1	Auswahl der zu bewertenden Endpunkte
TOP 1a)	Lungenfunktionsmessung (FEV_1, IC)
TOP 1b)	Verbrauch an Bedarfsmedikation
TOP 1c)	Kardiopulmonaler Belastungstest (CPET)
TOP 2	Bewertung der Relevanz von Effekten
TOP 2a)	Notwendigkeit der Bewertung der Relevanz
TOP 2b)	Standardisierte Mittelwertdifferenz (SMD) als Effektmaß
TOP 3	Verschiedenes / Verabschiedung

4.3 Protokoll der wissenschaftlichen Erörterung

- Datum: 16.06.2010, 14:00 bis 15:55 Uhr
- Ort: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Dillenburger Straße 27, 51105 Köln
- Moderation: PD Dr. Stefan Lange

4.3.1 Begrüßung und Einleitung

Stefan Lange: Ich darf Sie ganz herzlich zu der Erörterung der Stellungnahmen zum Berichtsplan A05-18 „Tiotropiumbromid bei COPD“ begrüßen.

Zu Beginn eine grundsätzliche Bemerkung: Die meisten von Ihnen werden schon einmal hier gewesen sein. Falls nicht, das Ganze wird stenografisch und elektronisch aufgezeichnet; das ist Ihnen auch mitgeteilt worden. Daraus wird ein Wortprotokoll erstellt. Sie haben sich im Grunde genommen schon mit diesem Verfahren einverstanden erklärt. Falls Sie es sich anders überlegt haben sollten, müssten Sie die Erörterung verlassen. – Das scheint nicht der Fall zu sein.

Eine weitere grundsätzliche Bemerkung – das wissen Sie aber auch: Wir wollen uns nicht über die grundlegende Methodik der evidenzbasierten Medizin unterhalten oder über das Auftragsverhältnis zwischen dem Gemeinsamen Bundesausschuss und IQWiG – das können wir hier leider nicht besprechen; das muss an anderer Stelle passieren –, sondern es wird im Wesentlichen um Punkte aus Ihren Stellungnahmen gehen, zu denen wir noch Nachfragen haben oder die für uns unklar geblieben sind.

Des Weiteren möchte ich Sie bitten, bei Wortmeldungen immer Ihren Namen zu nennen, damit der Stenografin das Mitschreiben leichter gemacht wird.

Ich muss mich noch vorstellen: Mein Name ist Lange, stellvertretender Leiter dieses Instituts. Ich werde die heutige Veranstaltung moderieren.

Gibt es dazu Ihrerseits Fragen? – Nein.

Letzter Punkt: Wie immer gibt es, wenn Sie noch etwas ganz Dringendes loswerden wollen, die Möglichkeit, uns das unter TOP 3 „Verschiedenes / Verabschiedung“ der vorbereiteten Tagesordnung mitzuteilen, die Ihnen vor einigen Tagen zugegangen ist, damit Sie Gelegenheit hatten, sich vorzubereiten.

Ich schlage vor: Beginnen wir jetzt!

Daniel Fleer: Ich begrüße Sie recht herzlich. Ich bin der Leiter des heute zu diskutierenden Projekts und werde die Diskussion bei TOP 1 leiten. Wir kommen zunächst zu:

4.3.2 Tagesordnungspunkt 1: Auswahl der zu bewertenden Endpunkte a) Lungenfunktionsmessung (FEV_1, IC)

Daniel Fleer: Im Berichtsplan wird die Lungenfunktionsmessung als zu bewertender Endpunkt nicht genannt. Von der Firma Novartis und der Deutschen Atemwegsliga wurde aber der Einschluss von Lungenfunktionsparametern als Surrogate für patientenrelevante Endpunkte angeregt. Konkret wurden die Einsekundenkapazität und die inspiratorische Kapazität genannt. Die eingereichte Literatur zeigte allerdings nur die prognostischen Eigenschaften dieser beiden Größen, und wir bitten Sie darzulegen, warum Sie davon ausgehen, dass eine Änderung dieser Werte durch eine Intervention auch einen konsistenten und gleichgerichteten Zusammenhang mit der Änderung von patientenrelevanten Endpunkten zeigt.

Stefan Lange: Genau, das ist ja die Definition von Surrogaten.

Peter Kardos: FEV_1 ist der am besten validierte Endpunkt für die Prognose. Es ist nicht optimal, aber wir haben nichts Besseres, und es ist in den meisten Studien vorhanden. Das ist schon ein Argument, FEV_1 aufzunehmen. Hinzu kommt, die Prognose, also die Lebenserwartung oder die Mortalität, ist ein absolut patientenrelevanter Endpunkt. Ich sehe, offen gesagt, keinen Grund für den Ausschluss. Es kommt noch hinzu, dass FEV_1 stellvertretend – wieder nicht optimal, aber immerhin – auch für solche Atemfunktionsparameter wie IC steht, was in den meisten Studien nicht gemessen wird. Insofern ist es illusorisch, das einzuschließen, weil wir dadurch ganz wenige Studien hätten. Aber wir wissen, dass es parallel geht, und wir wissen auch, dass dadurch Dyspnoe und Belastungsfähigkeit als interessante patientenrelevante Endpunkte mitberücksichtigt werden würden.

Lars Peters: Dem kann ich nur beipflichten. Die Studien, die wir mit angegeben haben, zeigen schon eine Korrelation zwischen FEV_1 oder auch inspiratorischer Kapazität mit Mortalität, aber auch mit weiteren patientenrelevanten Endpunkten wie Dyspnoe. Weiterhin, wie es auch im Berichtsplan steht, wird die Schweregradeinteilung der Patienten nun mal anhand dieser Werte vorgenommen. Wenn man sich jetzt vorstellt, dass bei einem Patienten mit einem höheren Schweregrad auch andere verschiedene patientenrelevante Endpunkte assoziiert sind, muss man den zwangsläufig mit hineinnehmen, um überhaupt messen zu können, wie der Krankheitsverlauf durch eine bestimmte Therapie beeinflusst wird.

Beate Wieseler: Sie haben den Stellenwert der Lungenfunktion, insbesondere des FEV_1, jetzt in zwei Bereichen dargestellt: zum einen als prognostischer Faktor für die Patienten und zum andern für eine Schweregradeinteilung, also im Bereich der Diagnose. Nun interessieren wir uns aber für Interventionseffekte, und da ist primär unsere Frage: Ist eine Änderung der Lungenfunktion als Reaktion auf eine Intervention konsistent gleichgerichtet in Relation zu patientenrelevanten Endpunkten zu sehen, zum Beispiel zur Symptomatik? Es gibt eine Reihe von Interventionsstudien, in denen das nicht der Fall zu sein scheint, wir also keine Änderung

in der Symptomatik, aber sehr wohl eine Änderung in der Lungenfunktion sehen und umgekehrt. Wir haben eigentlich in der Literatur Beispiele dafür gefunden, dass dieser konsistente Zusammenhang zwischen den Effekten von Interventionen auf die Lungenfunktion und auf patientenrelevante Endpunkte nicht besteht. Deshalb ist FEV_1 aus unserer Sicht kein valides Surrogat für patientenrelevante Endpunkte in dieser Indikation.

Peter Kardos: Es gibt sicher Studien, in denen eine schwache, aber positive und signifikante Korrelation von FEV_1 und Lebensqualität besteht – ein patientenrelevanter Endpunkt. In der Literaturstelle Ende der 90er-Jahre Paul Jones ist der Korrelationskoeffizient, glaube ich, 0,26 oder so, also schwach, aber immerhin. Es gibt mehrere Studien, die besagen, dass es eine enge Korrelation mit den Exazerbationen gibt. Die Exazerbationen sind ebenfalls ein absolut patientenrelevanter Endpunkt. Da sind die zwei Arbeiten von Frau Wedzicha im „Thorax“ zu erwähnen und eine gerade erschienene neue Arbeit. Dazu gibt es also etwas.

Stefan Lange: Herr Kardos, das ist unbenommen, das glauben wir ja auch. Der prognostische Wert ist wohl unbestritten. Aber noch mal: Es geht um die Frage, ob eine Intervention, wenn sie eine Veränderung im FEV_1 hervorruft, mit einer Veränderung in den patientenrelevanten Endpunkten gleichzusetzen ist. Nur dann ist es ein valides Surrogat. Es gibt klare Anforderungen an die Definition von Surrogaten. Hier sind auch ein paar Biometriker vertreten, die wissen: Ein Surrogat ist dann ein valides Surrogat, wenn der Effekt auf das Surrogat den Effekt auf den Endpunkt, für den man sich interessiert, erklärt. Das scheint nach den Äußerungen von Frau Wieseler nicht der Fall zu sein. Das wäre spannend, wenn Sie eine solche Literaturstelle hätten, also nicht die alleinige Korrelation. Das kennen wir in der Medizin aus anderen Bereichen zur Genüge, dass man da sehr schnell Irrtümern – fatalen übrigens – erliegen kann. Denken Sie etwa an das berühmte Cholesterin oder andere Beispiele!

Peter Kardos: Ich habe keine Literatur im Kopf, dass eine Änderung des Parameters FEV_1 mit einer Änderung … Aber das ist wahrscheinlich auch nicht gemessen worden. Das liegt einfach außerhalb des Fokus der Studien.

Beate Wieseler: Vielleicht kann ich auch noch darauf hinweisen, dass wir nicht alleine die Wertigkeit von FEV_1 als Outcome in Interventionsstudien bezweifeln. Auch eine Task-Force der amerikanischen und europäischen Fachgesellschaften zu Outcomes in COPD-Studien beschreibt eigentlich auch, dass Lungenfunktion schlecht korreliert ist mit patientenrelevanten Endpunkten wie Symptomatik und Exazerbationen. Ich denke nicht, dass wir da eine Außenseitermeinung vertreten, wenn wir davon ausgehen, dass die Lungenfunktion kein valides Surrogat in Interventionsstudien ist.

Peter Kardos: Also: schwache, aber positive Korrelation. Ich meine, im ATS/ESR des Jahres 2008, im Zweifel 2007, steht ja, dass das nicht, wie es bis dahin üblich war, als alleiniger Endpunkt benutzt werden kann. Dass es aber ganz eliminiert werden soll, ist mir so nicht geläufig.

Stefan Lange: Gibt es weitere Wortmeldungen dazu?

Lars Peters: Meinen Sie, wenn Sie von Lungenfunktion sprechen, nur FEV_1 oder auch inspiratorische Kapazität? Das haben wir in der Diskussion ein bisschen vermischt.

Beate Wieseler: Sehen Sie denn den Stellenwert der inspiratorischen Kapazität anders, und wenn ja, warum?

Lars Peters: Nun ja, es gibt wiederum Studien, die besonders zur Dyspnoe erklärend sind, dass die inspiratorische Kapazität ein Surrogat ist.

Stefan Lange: Können Sie uns die Studien nennen? Aber bitte in der Weise, wie ich es gerade erläutert habe: Der Effekt auf das Surrogat erklärt den Effekt auf den patientenrelevanten Endpunkt. Nur darum geht es; es geht nicht um die reine Korrelation von Surrogat und Endpunkt. Das haben wir in der Medizin hunderttausendfach. Das ist schön und auch die absolut notwendige Voraussetzung, aber sie ist nicht hinreichend, sondern wir benötigen den Nachweis, dass der Effekt auf das Surrogat den Effekt auf den patientenrelevanten Endpunkt erklärt. Das ist quasi eine Korrelation von Effekten. Dazu gibt es eine ausgereifte Methodik und in anderen Bereichen auch Meta-Analysen. Zum Beispiel ist in bestimmten onkologischen Indikationen jedenfalls teilweise das krankheitsfreie Überleben ein Surrogat in diesem Sinne für das Gesamtüberleben. Das gilt nicht für alle onkologischen Indikationen, aber für einige. Daran kann man sehr schön sehen, dass diese Effekte auf das Surrogat mit den Effekten auf den tatsächlichen Endpunkt korrelieren. Aber eine alleinige Korrelation zwischen Surrogat und Endpunkt ist für unsere Fragestellung völlig uninteressant.

Peter Kardos: Ich muss etwas fragen; ich weiß nicht, ob ich das richtig verstehe. Wenn in einer Studie die Anwendung eines Medikaments zu einer besseren körperlichen Belastbarkeit führt, dass also die körperliche Belastbarkeit a) länger ist und b) eine höhere Belastungsstufe erreicht werden kann, ist das in Ihrem Sinne für Sie von Interesse oder nicht?

Stefan Lange: Ja.

Peter Kardos: Es gibt drei, vier Studien von der kanadischen Gruppe – den Namen kann ich Ihnen gleich liefern; Magnussen aus Deutschland war Koautor –, in denen die inspiratorische Kapazität gemessen wurde und nachgewiesen wurde, dass die körperliche Belastbarkeit auf dem Ergometer länger ist und eine höhere Belastungsstufe erreicht werden kann. Ich sage Ihnen gleich den Autor.

Stefan Lange: Gut, Sie können ihn uns gleich noch nennen. Gegebenenfalls können Sie uns die Information auch über einen anderen Weg zukommen lassen. – Ich sehe keine weiteren Wortmeldungen.

Daniel Fleer: Wir kommen zu:

4.3.3 Tagesordnungspunkt 1: Auswahl der zu bewertenden Endpunkte b) Verbrauch an Bedarfsmedikation

Daniel Fleer: Die COPD-Symptome werden im Berichtsplan als zu bewertender Endpunkt genannt. Von der Firma Novartis wurde vorgeschlagen, den Verbrauch an Bedarfsmedikation, das heißt von kurzwirksamen Bronchodilatatoren, als Surrogat für die Kontrolle der Symptomatik explizit in den Berichtsplan als zu bewertenden Endpunkt aufzunehmen. Es wurde allerdings keine Literatur zur Begründung eingereicht. Deswegen meine Frage an Sie: Worauf stützen Sie Ihre Aussage, dass dieser Verbrauch so, wie Sie sagen, den Grad und die Häufigkeit von auftretenden Symptomen für den Patienten reflektiere, und wieso halten Sie dieses Maß für unverzichtbar in der Nutzenbewertung zusätzlich zur direkten Symptommessung?

Lars Peters: Ich muss Ihnen zunächst einmal recht geben, es ist auf jeden Fall ein Surrogatparameter, den man mit den anderen gemessenen Endpunkten – Lebensqualität, Dyspnoe – direkt messen kann. Auf der anderen Seite sollten wir das aber vielleicht nicht ganz hinten runterfallen lassen; denn die Angaben, die der Patient in den patient-reported Outcomes macht, werden mit allen Medikamenten, die er während der Studienzeit zu sich nimmt, gemacht. Dementsprechend sind wir der Meinung, dass es schon eine Rolle spielt, wie viele kurzwirksame Beta-2-Sympathomimetika er mit einnimmt. Denn die spielen mit in das Ergebnis hinein, wenn man insbesondere auf den SGRQ als Messinstrument zur Lebensqualität anspricht.

Beate Wieseler: Aber: Verstehe ich Sie dann richtig, dass Sie die Bedarfsmedikation also gar nicht als Surrogat für eine patientenrelevante Outcome-Symptomatik, sondern eher als eine Art Störgröße, als Effektmodifikator in der Studie, sehen würden?

Lars Peters: Dem kann ich zustimmen. Deswegen sollten die Ergebnisse – SGRQ, also die anderen Ergebnisse –, auch den Verbrauch an Bedarfsmedikation mit berücksichtigen. Denn wie Sie gerade gesagt haben, kann man davon ausgehen, dass der Verbrauch an Bedarfsmedikation diese Ergebnisse verzerren kann, wenn er in den verschiedenen Studiengruppen – Treatment Arms – sehr unterschiedlich ist.

Stefan Lange: Da haben Sie aber ein Problem. Wie wollen Sie das von einem echten Surrogat trennen? Wenn das ein echtes Surrogat wäre, würden Sie genau das erwarten. Insofern würden Sie sich damit jeglichen Effekt kaputt machen. Denn wenn Sie sagen, das liegt eigentlich nur an dem geringeren Verbrauch an Bedarfsmedikation oder umgekehrt – keine Ahnung –, dass sich Symptome verändern, wäre das genau die Erklärung für die Symptomänderung, und Sie würden mir sagen: Jetzt habe ich gar keinen Effekt mehr. – Das würde ich noch mal überdenken.

Lars Peters: Gut, es kann in manchen Fällen so sein. Wenn sich aber andersherum in diesem Beispiel die Werte SGRQ verbessern und gleichzeitig der Verbrauch an Bedarfsmedikation zurückgeht, kann man davon ausgehen, dass diese Resultate besonders robust sind.

Beate Wieseler: Nach wie vor meine Frage: Zielt Ihr Vorschlag jetzt eher darauf hin, das als Effektmodifikator oder als Surrogat für einen patientenrelevanten Endpunkt zu berücksichtigen?

Lars Peters: Mehr als Effektmodifikator.

Friedhelm Leverkus: Ich möchte den Hinweis geben, dass es natürlich in dem meta-analytischen Kontext nicht ganz einfach ist, solche Effektmodifikatoren irgendwo hineinzubringen, und durchaus eine gewisse Gefahr besteht, dass man Dinge oder Effekte wegadjustiert, die bereits Effekte sind, weil man nicht auf die Patientenebene heruntergehen kann.

Stefan Lange: Ich habe eine Nachfrage, Herr Leverkus. Würden Sie dann eher sagen, man sollte darauf verzichten?

Friedhelm Leverkus: Ich denke, dazu muss die Medizin etwas sagen, inwieweit das ein wesentlicher Faktor ist und inwieweit nicht noch andere Dinge eine Rolle spielen als nur die Gabe der Medikation. Aber ich bin bei solchen Dingen, gerade wenn sie während der Behandlung passieren – das ist eine Begleitmedikation, die nimmt er während der Behandlung; das ist nichts Unabhängiges; das kann auch der Effekt der Behandlung sein –, sehr, sehr vorsichtig. Es kann natürlich sein, dass man so etwas anwenden muss. Aber da muss man sehr, sehr vorsichtig sein, weil es nicht vor Randomisierung geschieht.

Stefan Lange: Da stimme ich Ihnen zu.

Harald Kögler: Ich möchte nur darauf hinweisen, dass der Stellenwert des Gebrauchs an Bedarfsmedikation bei der COPD nach unserer Einschätzung bei Weitem nicht derselbe ist wie beim Asthma. Dr. Kardos, Sie werden mir wahrscheinlich beipflichten. Beim Asthmakontrolltest, einem schnellen Beurteilungsverfahren für die klinische Praxis, ist der Verbrauch an Bedarfsmedikation als eines von fünf wesentlichen Kriterien berücksichtigt. Beim COPD Assessment Test, an dessen Entstehen Sie auch beteiligt waren, ist es nicht als eines von acht relevanten Kriterien selektiert worden. Auch in den Leitlinien, wenn man sich die deutsche Asthma-Leitlinie, die deutsche COPD-Leitlinie anschaut, ist der Verbrauch an Bedarfsmedikation beim Asthma wesentliches Element, um die Krankheitskontrolle zu bewerten, nicht jedoch bei der COPD. Von daher würde ich den Stellenwert insgesamt in dem Therapiegebiet, um das es sich bei der heutigen Anhörung handelt, als sehr gering bewerten.

Peter Kardos: Zu dieser Frage möchte ich sagen: Ich habe das Gefühl, statistisch wird es kaum möglich sein, diese Effekte auszuarbeiten und als patientenrelevantes Surrogat

nachzuweisen. Aber andererseits sage ich natürlich als Kliniker, gerade heute in der Diskussion um die potenziell tödlichen Nebenwirkungen von langwirksamen und kurzwirksamen Betamimetika ist das schon eine zumindest potenziell außerordentlich relevante Frage, wie viel Betamimetika der Patient inhaliert. Die Frage ist noch nicht entschieden, aber immerhin haben die Amerikaner mittlerweile die Black Box Warning auf den Packungen der langwirksamen Betamimetika: Dieses Medikament kann in seltenen Fällen zum Tode führen. – Dass die kurzwirksamen die gleichen Wirkungen haben, wissen wir seit 1967, der Arbeit von Inman. Da möchte ich bei dieser Gelegenheit noch die Arbeit mit der IC und Überblähung liefern. Das ist O'Donnell. Von O'Donnell gibt es mehrere Arbeiten; ich nenne jetzt „European Respiratory Journal" 2004.

Stefan Lange: Vielen Dank.

Beate Wieseler: Vielleicht noch eine Frage im Zusammenhang mit der Bedarfsmedikation, die den Endpunkt aufgreift, also den Unterschied zu Asthma. Ein wichtiger Punkt in der Unterscheidung der Diagnose von Asthma und COPD ist ja die Reversibilität. Ergibt sich nicht auch schon dadurch, dass ich bei COPD von einer sehr geringen Reversibilität ausgehe, dass die Bedarfsmedikation nicht den gleichen Stellenwert hat wie bei Asthma? Also: Welchen Stellenwert haben die kurzwirksamen Beta-2-Sympathomimetika bei COPD, bei der ich sowieso von einer sehr geringen Reversibilität ausgehe?

Peter Kardos: Sie haben im Prinzip recht. So steht es auch in den Lehrbüchern. Wenn man ein bisschen dahinterschaut und zum Beispiel FEV_1 als Endpunkt von großen Studien, hochkarätig publiziert, nimmt, werden Sie sehen, dass etwa die Peak-Flow-Änderungen in COPD-Studien und in Asthmastudien erstaunlicherweise von der Größenordnung sehr ähnlich sind. Nehmen wir an, zwischen 20 und 35 ml. Das wird als Riesenerfolg gefeiert. Manchmal ist es schwierig, das auseinanderzuhalten. Ich meine, die COPD-Patienten nehmen nicht weniger Betamimetika – Notfallmedikation oder Bedarfsmedikation – als die Asthmapatienten. Die sind komorbid, die sind häufiger komorbid als die Asthmapatienten. Klinisch gesehen besteht schon das Gefühl, ohne es mit Arbeiten belegen zu können – sehr schwierig –, dass das nicht unbedingt kardial und prognostisch günstig ist.

Norbert Banik: Dazu noch ein Punkt, um vielleicht die Brücke zu schlagen zwischen dem, was gesagt wurde, und dem, was Herr Leverkus noch eingewendet hat: Die bisherige Reaktion Ihrerseits verstehe ich so, dass die Anerkenntnis des Faktors „Verbrauch an Bedarfsmedikation" als effektmodifizierender Faktor nicht infrage steht. Aber es wäre sicher auch gut, im Berichtsplan und in der Beurteilung der Evidenzen daran zu denken.

Auf der andern Seite möchte ich als jemand, der auf diesem Gebiet auch selber Studien gestaltet, sagen: Es ist ganz unmittelbar so bekannt, wenn man Studien, die Symptomkontrolle als Endpunkt haben, gestaltet, dass man dann an diesen Faktor „Verbrauch an Bedarfsmedikation" denkt – entweder durch Kontrolle im Studiendesign oder durch Adjustierungsmaßnahmen. Und dann sollte man sozusagen auf Studienebene davon ausgehen,

dass diesem effektmodifizierenden Einfluss zu einem großen Anteil schon Rechnung getragen wurde. Insofern kann man wahrscheinlich in der Bewertung diesem Dilemma entgehen. Aber, wie gesagt, es wäre schön, wenn vielleicht im Berichtsplan eine Notiz dazu erschiene, dass Sie diesen Faktor in diesem Sinne auch mit bedenken.

Stefan Lange: Noch eine Nachfrage dazu?

Thomas Glaab: Ich denke, die Bedeutung der Bedarfsmedikation ist für Asthma und COPD klar. Wie Herr Kardos gesagt hat, selbst eine kleine Verbesserung des FEV_1 bringt dem Patienten – auch COPD-Patienten – durchaus einen individuellen Effekt, wenngleich wir wohl alle hier der Meinung sind, dass es sich um einen weichen Parameter handelt, der in Studien auch nie als primärer Endpunkt erfasst wird, bei dem die Korrelation zu klinisch relevanten Endpunkten auch nie so hergestellt wurde und auch in Erhebungen erhebliche Qualitätsmängel bestehen. Ich denke, so kann man das schon zusammenfassen.

Stefan Lange: Okay. Vielen Dank. – Dann kommen wir zu:

4.3.4 Tagesordnungspunkt 1: Auswahl der zu bewertenden Endpunkte c) Kardiopulmonaler Belastungstest (CPET)

Daniel Fleer: Der Berichtsplan nennt als zu bewertenden Endpunkt die körperliche Belastbarkeit. Die Firma Novartis hält es für notwendig, dass im Berichtsplan zu den konkreten Erhebungsinstrumenten für Belastbarkeit Stellung genommen wird. Genannt wird unter anderem der kardiopulmonale Belastungstest, also die Spiroergometrie. Bei dieser Untersuchung werden Patienten mittels eines Fahrradergometers oder eines Laufbandes körperlicher Belastung ausgesetzt und es werden verschiedene Parameter erhoben, zum Beispiel Ausdauer, blutgasanalytische Werte, ventilatorische Werte und kardiale Parameter. Welche von diesen bei dieser Untersuchung erhobenen Messgrößen sollten Ihrer Ansicht nach jetzt konkret in der Nutzenbewertung berücksichtigt werden?

Lars Peters: Um ehrlich zu sein, uns ging es mehr darum, überhaupt zu sehen, welche Belastungstests vom IQWiG anerkannt werden könnten. Wir haben zum einen den 6-Minute Walking Test genannt und die Tests, die Sie gerade vorher genannt haben – mehr als Beispiele. Für uns ist es, wie gesagt, einfach nur wichtig zu wissen, welche Formen der Belastungstests hier Anwendung finden können.

Beate Wieseler: Da sprechen Sie ein grundsätzliches methodisches Problem bei der Durchführung systematischer Übersichten an. Im Gegensatz zu einer prospektiven Planung einer einzelnen Studie, bei der ich mir natürlich vorher überlegen kann, welche Daten möchte ich erheben, in welcher Qualität und wie möchte ich die auswerten, ist das bei einer systematischen Übersicht, die eine retrospektive Studie darstellt, nicht möglich. Wir müssen im Grunde genommen dann mit den Daten arbeiten, die wir finden, und wir verzichten deshalb in der Regel darauf, einzelne Instrumente und Operationalisierungen im Berichtsplan zu nennen. Wir definieren vielmehr Kriterien und übergeordnete Endpunkte, die wir dann betrachten wollen, wenn auch nur, um nicht gegebenenfalls Dinge, die wir in der Form nicht präspezifiziert haben, ausschließen zu müssen.

Also: Es ist, denke ich, generell nicht sinnvoll, bei der Erstellung einer systematischen Übersicht alle denkbaren Operationalisierungen bestimmter Endpunkte zu benennen. Das ist der Grund, warum wir nicht schon einzelne Belastungstests beschreiben. Grundsätzlich sind alle Tests denkbar, die eine Belastbarkeit des Patienten zeigen können. Sie haben den 6-Minute Walking Test genannt. Das ist sicherlich so ein Instrument. Bei der Spiroergometrie wird sicherlich auch Belastbarkeit gemessen. Da ist vielleicht noch die Frage: Welche Parameter aus diesem Test sind die, die am besten geeignet sind, um tatsächlich die Belastbarkeit von COPD-Patienten abzubilden?

Stefan Lange: War die Antwort erst mal so weit befriedigend?

Lars Peters: Ja.

Stefan Lange: Okay, prima. Dann haben wir sozusagen Unklarheiten bei Ihnen beseitigen können. Das ist ja auch schön. – Möchte noch jemand etwas zu Belastungstests sagen?

Peter Kardos: Ich halte das für hochinteressant, und es gibt sicherlich kleinere Studien, in denen Sie die Sauerstoffaufnahme – mit und ohne Intervention –, die Belastungsdauer und die Belastungsintensität messen könnten. Ich bin skeptisch, dass eine solche Bewertung angesichts der Anzahl dieser Studien und der teilnehmenden Patienten bei der endgültigen Beurteilung ins Gewicht fallen würde. Das sind meine Bedenken. Ich bin Feuer und Flamme dafür. Wenn ich 20 Patienten untersuche, ist das hochinteressant und man kann wirklich Wirkungen nachweisen. Aber bei Ihrer Beurteilung bin ich skeptisch, ob Sie da weiterkämen.

Beate Wieseler: Sie würden einfach aus praktischen Erwägungen davon ausgehen, dass wir eher den 6-Minute Walking Test sehen als Spiroergometrie?

Peter Kardos: Ja. Der 6-Minute Walking Test bildet die Belastbarkeit oder die Fähigkeiten der Patientengruppe ab, die sozusagen die Zielgruppe für die Behandlung mit diesen Medikamenten sind. Denn das sind die, die mit dem Gehwagen noch 80 m oder 150 m laufen können, und unter Pharmakotherapie werden sie dann 180 oder 200 m laufen können, was Ihnen vielleicht erlaubt – entschuldigen Sie, dass ich das jetzt so sage –, alleine auf die Toilette zu gehen, was sie bislang nicht konnten. Das ist also ein patientenrelevanter Endpunkt. Aber, wie gesagt, man muss in der Literatur sehen, was dafür vorliegt, was Sie auswerten können. Relevant, ja.

Stefan Lange: Danke für diese Ergänzung. – Wir kommen zu:

4.3.5 Tagesordnungspunkt 2: Bewertung der Relevanz von Effekten

Guido Skipka: Bewertung der Relevanz ist ein Thema, das sich auf die Erhebung von Endpunkten mittels Fragebögen bzw. komplexer Skalen bezieht, wie das sicherlich bei der Bewertung von Bronchodilatatoren der Fall sein wird. Das Problem solcher erst mal künstlicher Skalen ist die Interpretation von Unterschieden, Differenzen. Es schließt sich sofort die Frage an, welche Größenordnung ein Effekt, sprich: ein Unterschied zwischen zwei Behandlungsalternativen haben muss, um als relevant eingestuft werden zu können.

Jetzt haben wir im Berichtsplan ein abgestuftes Vorgehen beschrieben. Wir möchten primär Responderanalysen, die auf validierten MIDs, sprich: Minimal Important Differences beruhen, heranziehen. Falls diese nicht zur Verfügung stehen, möchten wir auf die Ebene von Mittelwertdifferenzen zwischen den Gruppen gehen und schauen: Wenn es denn validierte Relevanzschwellen für dieses Distanzmaß gibt, werden wir diese benutzen. Gibt es diese nicht, möchten wir die Relevanzschwelle von 0,2 Standardabweichungen heranziehen, was letztendlich bedeuten würde, dass das Konfidenzintervall des Effektes komplett oberhalb von 0,2 sein müsste, gemessen an der standardisierten Mittelwertdifferenz. Das vorweg zur Einführung.

Diese vorgeschlagene Vorgehensweise wurde jetzt in einigen Stellungnahmen kritisiert. Wir haben TOP 2 in zwei Punkte untergliedert.

4.3.6 Tagesordnungspunkt 2: Bewertung der Relevanz von Effekten
a) Notwendigkeit der Bewertung der Relevanz

Guido Skipka: Es ist in einigen Stellungnahmen gesagt worden, dass bei der Betrachtung von patientenrelevanten Endpunkten doch jedes statistisch signifikante Ergebnis auch als relevant einzustufen sei, weil bei Betrachtung patientenrelevanter Endpunkte das Relevanzkriterium schon enthalten ist. Auf der anderen Seite wurde in einigen, zum Teil denselben Stellungnahmen – das ist wirklich interessant – gefordert, zur Frage von MIDs und Relevanzschwellen klinischen Sachverstand einzubinden.

Aus meiner Sicht ist es nicht so ganz vereinbar, wenn man auf der einen Seite sagt, allein die Beschränkung auf patientenrelevante Endpunkte ist ausreichend und auf der anderen Seite wird von Ihnen zum Teil gefordert, sich Gedanken über MIDs zu machen. Es stellt sich für uns schon die Frage, wie sich diese Sichtweisen vereinen lassen bzw. wie wir konkret vorgehen sollten.

Stefan Lange: Wer möchte hier den ersten Aufschlag machen? – Noch mal, die Frage ist: Besteht grundsätzlich Ihrerseits die Ansicht, wenn man patientenrelevante Endpunkte betrachtete, erübrigt sich jede Relevanzbetrachtung quantitativer Art? – Ich sehe niemanden, der aufzeigt.

Peter Kardos: Ich wundere mich, wahrscheinlich verstehe ich die Diskussion nicht. Natürlich ist zum Beispiel die Lebensqualität ein patientenrelevanter Endpunkt, etwa gemessen nach St. George's Respiratory Questionnaire. Eine Änderung, die statistisch hochsignifikant sein kann, ist manchmal oder in vielen Fällen, in vielen Studien so gering, dass der Patient das gar nicht merkt. Der Doktor muss also dem Patienten sagen: Es geht dir jetzt besser. – Das ist natürlich nichts wert. Insofern ist MCID – Sie sagen MID; das ist dasselbe – in meinen Augen schon unverzichtbar.

Stefan Lange: Okay. Jetzt geht es los. Herr Banik.

(Heiterkeit)

Norbert Banik: Das ist sozusagen genau die Brücke zwischen diesen beiden Argumenten. Da es angesprochen wurde, denke ich, kommt es nicht von ungefähr, dass man diese beiden Punkte auch diskutieren kann und soll. Denn Sie sagen selber, auch im Berichtsplan und bei der bisherigen Arbeit ist diese Hierarchie auch vom Institut selber angewendet worden, dass man sagt: Wenn es ein Responderkriterium oder eine bekannte MID gibt, dann wenden wir die an, und nur, wenn es die nicht gibt, dann nicht. Insofern ist es natürlich so, was Herr Kardos an dem Beispiel erläutert hat, dass, wenn eine Differenz zwar signifikant ist, aber es gibt eine bekannte MID, sagen wir eine inhaltlich ankerbasiert festgelegte, die damit nicht erreicht wird, dann würde die sozusagen die Relevanzbetrachtung des ohnehin patientenrelevanten Endpunktes überschreiben, also dominieren. Wenn es die aber nicht gibt,

dann ist genau das Feld für das Problem eröffnet, das wir hier diskutieren, nämlich: Was machen wir dann? Aber dass sozusagen beides gleichzeitig existiert – es gibt sie, und es gibt sie nicht –, das kann ja nicht vorkommen. Deswegen halte ich jemanden wie zum Beispiel mich, der das in der Stellungnahme unserer Firma geäußert hat, nicht von vorneherein für schizophren, weil das zwei getrennte und aus logischen Gründen immer getrennt sein müssende Entitäten sind.

Stefan Lange: Darf ich nur eines ergänzen, Herr Banik? Natürlich halten wir Sie nicht für schizophren, um das noch einmal protokollarisch festzuhalten,

(Heiterkeit)

aber Sie haben in Ihrer Stellungnahme tatsächlich geschrieben, dass Sie bei einem patientenrelevanten Endpunkt davon ausgehen, dass ein signifikantes Ergebnis per se relevant sei. Das haben Sie so formuliert. Von daher war das für uns gerade der Punkt. Gerade habe ich Sie aber ein bisschen anders verstanden, indem Sie sagen: Wenn wir all dies haben, dann orientieren wir uns auch an den MIDs und sehen da eine Relevanzbetrachtung schon für sinnvoll an.

Norbert Banik: Wenn ich direkt darauf antworten darf, stimme ich Ihnen schon zu. Ich würde es ergänzen – und so habe ich es auch gemeint: Per se ist alles, was signifikant ist, auch relevant, wenn es sich um einen relevanten Endpunkt handelt. Wenn aber bekannt ist, dass bestimmte Dinge für diesen Endpunkt schon als irrelevant nachgewiesen sind – wir brauchen zum Beispiel nur das Problem des Messfehlers ins Gespräch zu werfen, wenn also Änderungen unterhalb des Messfehlers liegen und durch eine große Fallzahl trotzdem zur Signifikanz geführt werden –, würden wir, würde ich auch zustimmen zu sagen, dann ist es, trotz dieser Prämisse, auch möglich, dass es nicht relevant ist. Aber per se ist die Diskussion damit erst mal auf einem anderen Niveau, finde ich, wenn man sagt: Patientenrelevante Endpunkte, die sich statistisch signifikant ändern, können relevant sein. Dann kann man Ausschlussgründe diskutieren, so vorhanden.

Stefan Lange: Nur, um es zu verstehen. Sie sagen: Wenn es eine MID gibt, Orientierung an MID. Wenn es keine gibt, ist jegliches signifikante Ergebnis als relevant zu betrachten. Habe ich Sie jetzt richtig verstanden?

Norbert Banik: Bei patientenrelevanten Endpunkten, ja.

Stefan Lange: Ja, natürlich bei patientenrelevanten Endpunkten. Das wäre allerdings ein Anreiz, möglichst keine MIDs für patientenrelevante Endpunkte auf Skalen zu etablieren. Das nur als Problem.

Norbert Banik: Das ist eine einseitige Sicht, die wir nicht vertreten.

Stefan Lange: Nein, das ist keine einseitige Sicht, sondern das ist nur die logische Konsequenz.

Norbert Banik: Das kommt darauf an, was man für ein Interesse hat.

Stefan Lange: Okay, gut.

Norbert Banik: Wir haben dieses Interesse nicht.

Stefan Lange: Okay.

Friedhelm Leverkus: Zuerst möchte ich sagen: Das ist ein schwieriges Thema, an das Sie herangehen. Dass Sie mehrere Stufen, Szenarien haben, finde ich im Prinzip gut. Generell denke ich auch, dass, wenn man einen patientenrelevanten Endpunkt oder gewisse patientenrelevante Endpunkte hat, jede signifikante Veränderung erst mal relevant ist. Nehmen wir zum Beispiel den Endpunkt Mortalität! Wenn ich zeigen kann, dass da ein signifikanter Unterschied ist, ist das erst mal relevant. Dann gibt es sicher einige Skalen wie Lebensqualität, bei denen man sich fragen muss: Was bedeutet das jetzt? Was ist überhaupt los? Da ist der Ansatz, wie es die Behörden machen, zu sagen, wir fordern ein statistisch signifikantes Ergebnis und gucken uns den Responder an, ein durchaus gangbarer Weg, um das zu verstehen. Das machen Sie im Prinzip auch. Der Punkt der Kritik geht vielmehr dahin, dass man jetzt versucht, die klinische Relevanz unabhängig von der Indikation zu sehen. Ja?

Stefan Lange: Zu dem Punkt kommen wir gleich unter b). Vielleicht können wir irgendwie versuchen, das zu trennen.

Friedhelm Leverkus: Okay.

Guido Skipka: Vielleicht noch zur Ergänzung, Herr Leverkus: Für Mortalität – das ist erst mal klar – streben wir keine Relevanzbewertung an. Wir beziehen uns wirklich auf komplexe Skalen bzw. Fragebögen. Vielleicht noch als Hinweis, aber das werden Sie ja wissen: Auch in den EMEA-Guidelines, den FDA-Guidelines zu Quality of Life und Patient-Reported Outcomes wird klar gesagt, dass die rein statistische Signifikanz bei diesen Endpunkten nicht ausreichend ist. Das ist also nicht unbedingt eine sehr exotische Meinung unsererseits.

Stefan Lange: Vielleicht darf ich noch zwei Dinge ergänzen:

Erstens. Das ist erst mal wichtig zu sagen, da gibt es noch Unterschiede zwischen zum Beispiel PROs und Mortalität. Nichtsdestotrotz, wenn es denn so ist, dass man sich bei Fragen der Mortalität nicht die Frage der klinischen Relevanz stellen darf – ich sage es jetzt mal überzogen –, dann dürfte es auch keine Nichtunterlegenheitsstudien in dem Bereich geben. Denn da muss man auch die Nichtunterlegenheitsgrenzen, also irrelevante Unterschiede, definieren. Da sehe ich einen gewissen Widerspruch. Denn in vielen Indikationsgebieten

werden auch Studien mit Endpunkt Mortalität – sogar alleine oder zumindest in Kombination – als Nichtunterlegenheitsstudien konzipiert.

Friedhelm Leverkus: Das ist sehr schwierig.

Stefan Lange: Es ist problematisch, wird aber gemacht, wohl auch von den Firmen, die Sie vertreten. Nur mal so am Rande, ich würde mir zutrauen zu behaupten, ich finde für jede Firma so eine Studie. Das ist jetzt so ein Faktencheck. Da können wir ja mal gucken, ob das hinhaut. Okay, das ist das eine.

Das Zweite, was, glaube ich, in dem Zusammenhang wichtig ist – das war mir bei Herrn Banik ein bisschen aufgefallen –, wir müssen noch mal versuchen, diese Ebenen zu trennen: MID – typischerweise für den Verlauf von Patienten in der Zeit –, um zu sehen, was ist denn jetzt für Patienten spürbar, gegebenenfalls als Responderkriterium heranziehen, und: Was ist ein relevanter Gruppenunterschied? Das sind unter Umständen zwei verschiedene Paar Schuhe. Nur, dass wir mal versuchen zu trennen.

Friedhelm Leverkus: Da gebe ich Ihnen vollkommen recht, das ist die Krux, um die es eigentlich geht. Es ist relativ einfach oder es geht sicherlich aus klinischer Sicht – da werden die klinischen Kollegen vielleicht noch etwas dazu sagen können –, für einen einzelnen Patienten Erfolgskriterien zu machen, obwohl das sicherlich auch variieren kann. Aber wenn man versucht, das auf die Gruppe hochzuziehen – Victor hat darüber schon in den 80ern geschrieben –, ist das nicht das Gleiche. Zum andern bedeutet das sehr, sehr viele Werturteile, zum Beispiel über Mortalität, wenn wir etwa sagen: Okay, da habe ich in den beiden Cancer-Gruppen nur einen Unterschied von 1 % Mortalität; das ist für mich jetzt nicht relevant. Das ist ein Werturteil, das man wissenschaftlich nicht belegen kann, sondern bei dem irgendjemand sagen muss: Das ist es mir nicht wert. Da sind wir meines Erachtens auch eher bei Fragen der Kosten-Nutzen-Bewertung: Was bin ich bereit als GKV oder als Gesellschaft für solche Unterschiede zu bezahlen?

Stefan Lange: Okay. Ich glaube, es ist wichtig, jetzt zu versuchen, die Diskussion zu begrenzen, weil wir nicht über binäre Endpunkte bis hin zur Mortalität sprechen, sondern wir gehen erst mal zu den komplexen Skalen. Fangen wir sozusagen klein an! Da hatte ich Herrn Kardos und auch Herrn Banik so verstanden: Es wird wohl Unterschiede geben – auch als Mittelwertunterschiede zwischen Gruppen –, die wir eigentlich nicht als relevant bezeichnen können. Habe ich Sie richtig verstanden?

Peter Kardos: Mich ja.

Stefan Lange: Jetzt kommt die nächste Frage. Von daher sind wir uns bei dem ersten Punkt, glaube ich, einig. Man kann auch nicht an den Guidelines vorbei. Das haben schon die FDA, die EMA – oder wie sie jetzt heißt – und auch spezifische Leitlinien immer wieder gesagt:

Neben der statistischen Signifikanz möge man auch die klinische Relevanz bewerten. Und wir haben eine qualitative Komponente; wir haben aber auch eine quantitative Komponente.

Jetzt ist vielleicht der nächste Punkt: Woran orientiert man sich? An dem, was man in der Studie als Punktschätzer sieht: Ist das kleiner als irgendetwas, was man meint, was relevant sei, oder kann es noch andere Kriterien geben? Das wird die spannende Frage sein, glaube ich. Ich habe Sie so verstanden, Herr Banik, dass Sie sagen, Sie gehen tatsächlich zum Punktschätzer hin. Sie haben, glaube ich, auch das Papier Kieser / Hauschke zitiert.

(Norbert Banik: Ja, da sind wir bei Punkt b)!)

Sind wir schon bei Punkt b)?

Guido Skipka: Nein, das ist eigentlich ein noch anderer Punkt.

Stefan Lange: Okay, dann ziehe ich das wieder zurück; das ist zwischen a) und b).

Guido Skipka: Das ist eigentlich kein Punkt aus Ihren Stellungnahmen gewesen, sondern aus der Stellungnahme von nächster Woche.

(Heiterkeit)

Stefan Lange: Ach so, okay.

(Norbert Banik: Das stimmt nicht; das ist schon ein Punkt der Stellungnahme! – Friedhelm Leverkus: Das haben wir mit drin in der Stellungnahme!)

Es ist doch mit drin.

(Norbert Banik: Ja, das ist auf alle Fälle mit drin!)

Gut, darüber können wir nachher noch sprechen. Das machen wir gleich bei Punkt b).

Harald Kögler: Ich würde doch noch mal eine Lanze für das Responderkriterium brechen wollen. Ich würde weggehen vom Punktschätzer zum Responderkriterium. Ein Patient, der in einer Studie ist, hat ja nicht die Alternativerfahrung, in den anderen Arm randomisiert worden zu sein. Also: Der Parallelgruppenvergleich ist für den Patienten – wir reden ja hier über Patientenrelevanz – nicht erheblich. Für ihn geht es darum: Wie ging es mir zu Beginn der Studie und wie geht es mir jetzt, zu dem Zeitpunkt, zu dem die Abschlussevaluation erfolgt, im Vergleich zu vorher? Habe ich mich subjektiv verbessert oder – das wäre natürlich auch wichtig – wenigstens nicht verschlechtert? Diese Dinge kann man, denke ich, wenn man die MIDs, wenn sie im Fachgebiet validiert sind, als Kriterium für patientenindividuelles Ansprechen festlegt, eigentlich doch recht gut fassen. Ich halte das für ein sehr geeignetes Vorgehen.

Stefan Lange: Das entspricht dann sozusagen auch unserer eigenen Idee, zu sagen: Also, das nehmen wir mal als Erstes. Wenn das vorliegt, prima, dann ziehen wir gerne solche Responderanalysen heran und würden dann auch tatsächlich jeglichen klinischen signifikanten Unterschied als relevant erachten. Im Grunde genommen ist das so ein bisschen das, was Sie sagen, Herr Banik. Okay? Aber wenn wir es nicht haben, haben wir ein Problem.

Guido Skipka: Kommen wir zu:

4.3.7 Tagesordnungspunkt 2: Bewertung der Relevanz von Effekten b) Standardisierte Mittelwertdifferenz (SMD) als Effektmaß

Guido Skipka: Kommen wir zu dem Fall, wenn wir diese Responderanalysen nicht haben! Faktisch läuft das darauf hinaus, dass wir die standardisierte Mittelwertdifferenz betrachten. Das wurde in Ihren Stellungnahmen kritisiert, da diese standardisierten Mittelwertdifferenzen rein verteilungsbasierte Verfahren sind, denen der klinische Anker fehlt. Das ist sicherlich auch unbestritten, dass dieser klinische Anker bei der SMD fehlt.

Dem würde ich als Biometriker etwas provokativ entgegenhalten, dass wir natürlich in statistischen Auswertungen, in klinischen Studien eine ganze Reihe von Festlegungen treffen, die keinen klinischen Anker haben: mal ganz banal angefangen beim Signifikanzniveau, das wir universell, ungeachtet spezieller klinischer Situationen, immer auf 5 % festlegen, oder die vorhin schon angesprochenen Nichtunterlegenheitsstudien. Ich kenne zumindest ein paar und weiß, dass es welche gibt, bei denen die Nichtunterlegenheitsgrenze auch ohne klinischen Anker festgelegt worden ist.

Noch mehr auf die Basis zurück: Unser ganzes Prinzip des statistischen Testens und eines statistisch signifikanten Ergebnisses beruht eigentlich einzig und allein auf einer standardisierten Mittelwertdifferenz. Man muss sich nur die Teststatistik vom t-Test mal ansehen; das ist die SMD. Bei einem statistisch signifikanten Ergebnis, aus dem Folgerungen gezogen werden, spielen erst mal klinische Aspekte keine Rolle. Die Signifikanz kann man immer aussprechen.

Das als Bemerkung dazu, jetzt ist unsere konkrete Frage: Was sollen wir anstelle der SMD benutzen, wenn wir keine Responderanalysen haben?

Friedhelm Leverkus: Ich wollte noch etwas zum Testproblem, zum klinischen Anker sagen. Ich denke, es ist ein grundsätzlicher Unterschied, ob ich ein Testproblem habe, bei dem es darum geht auszuschließen, dass irgendetwas größer als null ist, eine probabilistische Aussage, oder darum, Daten zu interpretieren. Das ist das, was diese klinische Relevanzproblematik eigentlich sagt. Auch wenn man noch mal in die Guidelines hineinschaut, geht beispielsweise EMEA von der Skala aus und möchte eine Interpretation der Daten haben, damit es der EMEA leichter wird, und schaut sich deshalb die Responderraten an. Denn in den EMEA-Guidelines ist noch einmal zitiert, dass man, wenn man über klinische Relevanzgrenzen geht, keine standardisierten Mittelwertdifferenzen nehmen, sondern versuchen soll, einen Konsens oder so etwas zu erzielen. Jeder, der bei der Planung einer Non-Inferiority-Studie mitgewirkt hat, weiß, wie schwer es ist, irgendwie solche Grenzen zu finden und zu einem Konsens zu kommen, um irgendwann dagegen zu testen. Von daher ist es meines Erachtens vielleicht sinnvoller zu versuchen, vorab einen Konsens zu machen. Das kann man in einer Art Delphi machen, im Scoping oder mit irgendetwas, was weiß ich.

(Stefan Lange: Immer dieses böse Wort!)

Man kann es anders nennen. Man kann es eine Art Delphi nennen, wo ein gewisser Konsens da ist, das aber nicht so zu nehmen, dass man dagegen testet, sondern dass das einem hilft, das zu interpretieren. Denn wenn ich dagegen teste, habe ich wieder eine Null-eins-Entscheidung: signifikant – nicht signifikant. Das wird der Unsicherheit, die wir letztlich haben, um die klinische Relevanzgrenze festzulegen, nicht gerecht. Die Entscheidung „null“ oder „nicht null“ oder „größer als null“ können wir irgendwie treffen, aber das klinischRelevante ist sehr, sehr viel ein Werturteil.

Stefan Lange: Herr Banik, wollen Sie noch etwas ergänzen?

Norbert Banik: Ich will versuchen, darauf zu antworten, um die Basis für die weitere Diskussion zu finden, was wir denn nun machen können und über was wir noch zu diskutieren haben.

Zunächst direkt zu Herrn Skipka: Ich denke, der generelle Verweis auf bestimmte Annahmen, die man in dem Feld nun mal hat, am Beispiel des Signifikanztests ist sicher richtig. Nun wird es aber nicht besser dadurch, indem man das Problem sozusagen multipliziert und auf den Signifikanztest mit nicht verschobener Nullhypothese noch mal einen Signifikanztest mit verschobener Nullhypothese draufsetzt. Wenn man schon anerkennt, dass bestimmte Setzungen fraglich sind und vielleicht nicht universell gelten, wird es nicht besser, wenn man jetzt über Antidepressiva und COPD-Lebensqualität immer den gleichen Leisten schert. Das wäre die Eingangsbemerkung zu diesen Setzungen.

Das Zweite ist, wenn man nun sagt, wir erkennen an, dass verteilungsbasierte Maße wie die standardisierte Mittelwertdifferenz im Grunde genommen, wie das IQWiG selber schon gesagt hat, ein Surrogat für diese Relevanzbeurteilung sind, weil wir eben in bestimmten Situationen nichts Besseres haben, ergibt sich immer noch die Folgefrage, wie man es operationalisiert. Wir haben schon gesagt, wir haben das zusätzliche Problem von individuellen Änderungen zu Gruppenänderungen. Wir haben also viele Probleme, die ungelöst sind.

Dann sollte man sich auch noch mal vergegenwärtigen, dass es im Kontext der Nutzenbewertung vielleicht doch mehr um eine Gesamtbeurteilung der vorliegenden Evidenz geht als darum, wirklich auf den Punkt zu sagen: Da ist irgendetwas soundso, und dann will ich es gar nicht mehr sehen und gar nicht mehr einbeziehen. Das ist das, was leider im Zusammenblick mit der späteren Kosten-Nutzen-Bewertung passiert. Auch diese Konsequenz muss man sich dann noch vor Augen führen: Wenn das Go oder No-go ist zu sagen, habe ich da einen Hinweis auf einen Nutzen oder einen Nachweis eines Nutzens, dann ist das schon sehr entscheidend. Da ist die Frage, wann ich diesen Cut-off „Das kommt in die Tonne; das will ich gar nicht mehr sehen“ setze, oder: Wann halte ich mir das für die Gesamtbetrachtung der Nutzenaspekte noch warm?

Nach dieser langen Rede – pardon – bliebe übrig, noch mal über die Operationalisierung zu reden, weil mir das wichtig wäre und es auch ein großer Teil der Stellungnahme war. Denn da spielen eben die Dinge hinein, die tatsächlich auch schon angeklungen sind – Analogie zu Nichtunterlegenheitsstudien usw. –, um zu sagen: Was ist denn dann eine mögliche Relevanzschwelle für diesen Test der verschobenen Nullhypothese, wenn man es schon so machen muss?

Stefan Lange: Okay. Bis jetzt habe ich noch keine Antwort auf unsere Frage gehört, wie wir es denn anders machen sollten.

(Norbert Banik: Das kommt ja noch!)

Okay, da freue ich mich. – Zu den vielen Punkten, die jetzt genannt sind, kann man einiges sagen.

Guido Skipka: Ich wollte noch mal grundsätzlich anmerken: Das ist alles richtig, was hier zu dem Punkt gesagt wird. Ich bitte Sie, nicht zu vergessen, dass die Betrachtung der standardisierten Mittelwertdifferenz und dieser Schwelle 0,2 unser letztes Mittel ist, um uns dieser Relevanzbewertung zu nähern. So, wie ich TOP 2a) und die Diskussion verstanden habe, sträubt sich niemand vehement dagegen, dass wir bei solchen Skalen eine Relevanzbewertung machen müssen. Und wenn wir auf dieser dritten Ebene der SMD-Betrachtung sind, wäre die Alternative, wenn wir das so nicht machen, es gar nicht bewerten zu können. Es könnte durchaus sein, dass eine Studie es trotzdem nach diesem vielleicht unsinnigen, vielleicht auch zu strengen Kriterium zeigt. Das wäre dann auch nicht so schlecht.

Stefan Lange: Oder eine Meta-Analyse von Studien.

Guido Skipka: Gut, das sowieso. Dass Meta-Analysen noch mal eine höhere Power haben, das ist Ihnen ja klar.

Beate Wieseler: Ich möchte noch einen Punkt von Herrn Leverkus aufgreifen. Sie sagten, es geht primär um eine Interpretation dieser Daten. Ich glaube, es geht um mehr. Es geht genauso wie bei den anderen Endpunkten darum, im Endeffekt eine Entscheidung zu treffen, ob ein Nutzen vorliegt oder nicht, und es geht auch darum, diese Entscheidung mit einer ausreichenden Sicherheit zu treffen.

(Friedhelm Leverkus: Okay!)

Das vielleicht auch noch als Hintergrund zu dem Verfahren, das wir hier gewählt haben.

Friedhelm Leverkus: Und das ist das Problem, die Sicherheit haben Sie eigentlich nicht, sondern Sie legen ein δ fest, das im Prinzip hocharbiträr ist, und nehmen dann etwas Sicheres, einen statistischen Test, und sagen: Batsch, ich habe die Entscheidung. Das ist meines Erachtens zu einfach, weil das unsicher ist. Es kann sein, dass in einer Situation, wo das δ von

0,2, die Standarddifferenz, klinisch irrelevant ist, aber es mag Situationen geben, wo es klinisch relevant ist. Von daher müsste man zum Ersten überlegen: Wie sieht das in der Situation aus?

Zum Zweiten ist das auch eine Wertbeurteilung. Man kann das statistisch nicht irgendwo festlegen, sondern man muss sagen: Okay, da ist es vielleicht ein bisschen kleiner, da ein bisschen größer.

Stefan Lange: Das ist völlig richtig.

Beate Wieseler: Ja, aber wir sind, wie Herr Skipka gerade noch mal ausgeführt hat, tatsächlich in der Situation, das als letzten Ausweg zu betrachten. Und wenn es Situationen gibt, in denen gegebenenfalls diese kleinen Differenzen auch relevant sind, würde ich Sie bitten, das für diesen Parameter in dieser Population zu zeigen. Da würde ich die Beweispflicht bei Ihnen sehen.

Stefan Lange: Um es noch mal zu konkretisieren: Wir planen ein hierarchisches Vorgehen. Gibt es eine MID? Ist die validiert? Wenn ja: Gibt es Responderanalysen? Falls es die gibt, ziehen wir die Responderanalysen heran und enthalten uns jeglichen Verschiebens einer Nullhypothese. Da haben wir Ihren Punkt schon aufgegriffen.

Okay, es gibt keine Responderanalysen, aber es gibt eine validierte MID. Dann sagen wir, wir werden diese validierte MID in die Relevanzbewertung, in die Festsetzung einer Relevanzschwelle einbeziehen. Das ist nicht 0,2, sondern das hängt von der MID ab. Wir haben am Freitag ein Symposium oder „IQWiG im Dialog“ – so nennen wir das – zu der Frage. Vielleicht können wir da auch darüber sprechen.

Das Dritte ist, wenn auch keine validierte MID da ist: Was sollen wir dann machen? Sie haben gesagt: ein Delphi. Okay. – Aber wir haben jetzt gesagt: 0,2.

Friedhelm Leverkus: Nehmen wir mal an, Sie nehmen die 0,2! Dann können Sie sich die 0,2 anschauen, etwas von dem sagen, was Sie wissen können, und versuchen, das zu bewerten. Aber meines Erachtens sollten Sie nicht testen, ob das Konfidenzintervall über den 0,2 ist. Wenn Sie eine qualitative Einschätzung geben und sagen, wir haben uns umgehört, haben mit den Leuten gesprochen, haben dies und jenes gemacht, vielleicht auch mal mit einer Normalverteilung versucht, Response Rates abzuschätzen, und kommen zu dem Urteil, dass es möglicherweise ein bisschen schwieriger ist, hat dieses Urteil eine andere Qualität als die Aussage: Der Nutzen ist nicht nachgewiesen.

Stefan Lange: Noch mal – das ist wichtig –, damit es klar wird:

Erstens. Trotzdem kommen wir irgendwann in die Situation – das hat Frau Wieseler schon gesagt –, entscheiden zu müssen, nicht nur wir, sondern alle. Da kommt man nie drum herum.

Man kann nicht immer sagen: Na ja, das muss man erst mal noch bewerten. Irgendwann muss man entscheiden; irgendwann ist die Welt binär.

(Friedhelm Leverkus: Ja, irgendwann schon!)

Genau. – Aber das Zweite ist, ich bin ein bisschen erstaunt, dass Sie sagen: Ja, gucken Sie doch mal so, und dann machen wir es vielleicht so, und dann beziehen wir noch das ein. – Sonst fordern Sie uns in Stellungnahmen immer auf, klare Kriterien zu benennen, damit Sie auch Sicherheit haben. Jetzt benennen wir klare Kriterien, das ist aber auch doof. Da weiß ich, ehrlich gesagt, nicht mehr, was wir noch machen sollen.

Peter Kardos: Ich wollte nur eine Frage stellen: Wie viele Endpunkte, die Sie ja einbeziehen wollen, sind hierdurch betroffen? Denn für die meisten patientenrelevanten Endpunkte gibt es MCIDs. Insofern wollte ich mal die Größenordnung sehen, über die wir hier diskutieren.

Beate Wieseler: Die Frage wird dann auch wieder sein, selbst wenn es diese MCIDs gibt: Haben wir Responderanalysen? Ansonsten werden wir auch wieder mit diesem Problem konfrontiert werden.

Norbert Banik: Man könnte vielleicht zu Ihrem Beitrag auch noch mal sagen, dass es wirklich so ist, wenn wir in dieser Hierarchie schon so weit runter sind, könnte man sagen: Gut, das ist es jetzt. Wir beschreiben das, was wir gesehen haben, weil wir anerkannt haben, dass alles, was wir dann noch machen würden, ein Werturteil erfordert und einbezieht. Dann kommt ja – das darf man nicht völlig vergessen – die nachgeschaltete Kosten-Nutzen-Bewertung, die im Grunde genommen über die Kosten, die die Gesellschaft bereit ist, dafür zu zahlen, entscheidet. Und – das darf ich noch sagen – diese Entscheidung treffen zum Glück nicht Sie, sondern der Gemeinsame Bundesausschuss. Die müssen dann sagen: Das ist uns das, was es kostet, wert – unter Umständen mag es ja gar nichts kosten; das wissen wir hier am Tisch nicht – oder es ist uns das nicht wert. Da ist doch dann die Wertentscheidung angebracht. Wenn Sie sozusagen den Scharfrichter spielen und sagen, die untere Grenze des Konfidenzintervalls war 0,19; da gibt es keinen Nutzen,

(Stefan Lange: Keinen Nutzenbeleg!)

keinen Nutzenbeleg, dann finden wir das zu scharf, weil wir auch noch ein bisschen über die Operationalisierung reden wollen. Das noch mal zu diesem Zusammenhang zwischen Werturteil und dieser Hierarchie, wo man zugegebenermaßen schon sehr weit unten ist.

Stefan Lange: Bitte die Kosten-Nutzen-Bewertung außen vor lassen! Das ist heute nicht Thema. Wir reden über Nutzenbewertung. Ich sage Ihnen auch direkt, warum. Sie brauchen den Kunstgriff auf die Kosten-Nutzen-Bewertung gar nicht. Wir haben das SGB V, in dem klar steht bzw. in Begründungen zum SGB V, aber auch in Vorschlägen für die Änderung, dieses AMNOG, dass es auch Aufgabe der Nutzenbewertung ist, die Patientengruppen, die

einen besonderen Nutzen haben, festzustellen. Das beschreibt wohl das Ausmaß des Nutzens. Wir kommen um diese Frage nicht herum, unabhängig davon, ob es um Kosten-Nutzen-Bewertung oder irgendetwas anderes geht. Wir müssen Patientengruppen beschreiben, die einen besonderen Nutzen haben. Wie würden Sie das denn anders verstehen, als dass es darum geht, eine Relevanzbewertung anzustellen?

Norbert Banik: Ich höre immer wieder, dass Sie sagen: beschreiben.

Stefan Lange: Ja.

Norbert Banik: Dann beschreiben Sie es doch!

Stefan Lange: Dann gibt es eine Entscheidung. Da ist ein besonderes Ausmaß des Nutzens da oder nicht.

Friedhelm Leverkus: Patientengruppen mit besonderem Nutzen verstehe ich eher als Subgruppen, dass man irgendwo Subgruppen hat.

(Stefan Lange: Ja!)

Aber das ist etwas anderes als die Entscheidung, die Sie im letzten Schritt machen wollen. Ich würde mich dem Norbert anschließen: Ein Medikament – Sie haben wirklich ein gutes Konzept mit den Effizienzgrenzen ausgearbeitet –, das nur halb so viel wie ein anderes Medikament bringt, darf im Prinzip auch nur halb so viel kosten. Das wäre eine Lösung, die …

Stefan Lange: Okay. Lassen Sie bitte die Kosten-Nutzen-Bewertung außen vor!

Guido Skipka: Ich habe es noch nicht ganz verstanden. Ich möchte jetzt auch mal Kosten-Nutzen-Daten außer Acht lassen. Sie sagen, wir sollen es einfach beschreiben. Meinen Sie das so, dass wir letztendlich das Konfidenzintervall, das die Meta-Analyse im Idealfall liefert, abbilden, oder meinen Sie noch ein bisschen mehr, oder gehen Sie auf diesen Kieser-Hauschke-Vorschlag, den Sie ja kennen? Dazu würde ich gerne noch ein bisschen was von Ihnen hören, wenn es möglich ist.

(Norbert Banik: Gerne!)

Einfach beschreiben – alles, was wir machen, ist irgendwie auch Beschreibung.

Stefan Lange: Nein. Das war die Frage b): Was würden Sie uns vorschlagen? Noch mal: Wir müssen eine Empfehlung zur Beschlussfassung geben. Das heißt im Grunde genommen, wir müssen schon den Gemeinsamen Bundesausschuss ziemlich treten und sagen: Wenn wir entscheiden würden, würden wir so entscheiden. Ihr könnt noch irgendwelche anderen Dinge heranziehen. – Deswegen müssen wir irgendwann zu Potte kommen und können nicht sagen,

so oder so, sondern null / eins. Das ist unser Auftrag. Da kommen wir nicht drum herum. Also: Wie können wir diesen Auftrag erledigen?

Norbert Banik: Ich werde gerne immer wieder danach gefragt, wie ich es denn machen würde, obwohl ich es in mindestens zwei Stellungnahmen im Abstand von einem Dreivierteljahr jeweils vorgeschlagen habe. Im Grunde genommen kommen beide Vorschläge zu der gleichen Quintessenz. Vor einem Dreivierteljahr habe ich es mit Kieser / Hauschke begründet. Inzwischen habe auch ich ein bisschen weiter nachgedacht und komme noch auf andere Begründungen und sage im Zusammenhang mit der dann ja doch irgendwie nachgeschalteten Kosten-Nutzen-Bewertung: Es wäre mir genug, wenn der Punktschätzer der standardisierten Mittelwertdifferenz größer oder gleich der verteilungsbasiert angenommenen Relevanzschwelle ist. Ob sie immer 0,2 sein sollte, würde ich mit Nein beantworten. Aber sagen wir mal, im Kontext COPD, Tiotropium ist sie vielleicht 0,2. Dann wäre dem so.

Stefan Lange: Ist das die MID? In Ihrer Stellungnahme haben Sie auch zwei δ.

Norbert Banik: Das ist noch gar nichts von den beiden. Ich rede jetzt über den Punktschätzer, gemessen an Ihrer Entscheidungsschwelle, und das ist für Sie – das wäre jetzt zu diskutieren – die Schwelle, ab der es sicher relevant ist. So habe ich das immer verstanden.

Stefan Lange: Genau, okay. Das würde ich dann MID nennen.

Norbert Banik: Ja, ach so. Und die notwendige Konsequenz dessen ist eigentlich, dass man sagt: Gut, wenn diese Schwelle – sagen wir 0,2, Cohen's d vor Kurzem, Hedges' g heute – die Schwelle ist, ab der es sicher relevant ist, dann wäre aber doch die logische Konsequenz, dass es unter dieser Schwelle nicht sofort irrelevant wird, sondern es gibt einzelne Patienten, Patientenuntergruppen, die da ebenfalls noch relevante Änderungen haben.

Wenn Sie da auch noch nicken, würde ich sagen, dann muss doch eigentlich jeder, der sich mit Biometrie und dieser Methodik beschäftigt hat, zu dem Schluss kommen: Aha, dann wäre doch im Grunde genommen in Analogie zur Nichtunterlegenheitsstudie der nächste Schritt, den statistischen Signifikanztest mit der verschobenen Nullhypothese nicht gegen die Schwelle, die sicher relevant ist, zu machen, sondern gegen eine Irrelevanzschwelle. Und diese Irrelevanzschwelle kann im Allgemeinen kleiner oder gleich dieser anderen Schwelle sein.

Dann sind wir bei der Diskussion: Wo ist diese Schwelle wirklich? Aber dass sie immer gleich dieser sicheren Relevanzschwelle ist, das würde ich nicht akzeptieren. Meine Argumentation, um das jetzt abzuschließen – mit ein bisschen anderen Argumenten früher bei den besonderen Antidepressiva und jetzt hier –, ist, dass diese Irrelevanzschwelle im besonderen Kontext der Nutzenbewertung in Deutschland mit der nachgeschalteten Kosten-Nutzen-Bewertung bei patientenrelevanten Endpunkten null sein sollte. Das führt operativ auf

den gleichen Vorschlag: Punktschätzer mindestens so groß, das Ganze statistisch signifikant, aber weitere Kriterien werden nicht herangezogen, werden der Interpretation anheimgestellt.

Stefan Lange: Okay. Ich präzisiere es nur. Also, Sie sagen, wir haben im Grunde genommen zwei Grenzen: sichere Relevanz, sichere Irrelevanz. – Völlig d'accord. Das können wir schon mal für das Protokoll festhalten.

Zweitens sagen Sie: Bei patientenrelevanten Endpunkten gibt es nichts sicher Irrelevantes. – Das haben wir vorher besprochen und festgestellt: Das ist zumindest verwirrend, es gibt einen gewissen Dissens. Ich halte es für nicht richtig; ich hatte das Beispiel Nichtunterlegenheitsstudien. Da haben Sie genau diese nicht sicher irrelevante Differenz. Die müssen wir immer definieren.

Zum Dritten stellen Sie fest: Es wäre günstiger, eine Relevanzbetrachtung am Punktschätzer zu machen. Aber dann, bitte schön, für das sicher Relevante.

Guido Skipka: Herr Banik, in Ihren letzten Ausführungen waren wir so ein bisschen auf demselben Weg, und dann sind Sie am Ende doch wieder abgebogen und wieder auf die Null gegangen. Das finde ich ein bisschen schade. Aber nichtsdestotrotz.

(Norbert Banik: Das kann ich auch erklären!)

Ich habe es aus Ihrer sehr ausführlichen Stellungnahme so herausgelesen, dass Sie argumentieren, unsere 0,2, das ist die Grenze, bei der wir sagen: Alles darüber ist sicher relevant, entspricht also in der Symbolsprache diesem δ_{rel}. Ich habe das bis jetzt bei uns immer anders gesehen. Die 0,2 sehe ich eher als das δ_{min}. Das δ_{rel} ist für mich noch größer als 0,2.

(Stefan Lange: In jedem Fall!)

Es gibt viele Bereiche in der Medizin, in denen man mit 0,5-Standardabweichungen arbeitet. Das nur zur Erklärung. In meinem Kopf sind die 0,2 dieser kleineren Grenze zuzuschreiben.

Norbert Banik: Da sind Sie aber nun im Konflikt mit Herrn Lange, der die 0,2 vorher schon abgenickt hat.

Stefan Lange: Nein.

Guido Skipka: Ich glaube, er hat nur genickt, weil er Sie verstanden hat.

Stefan Lange: Genau, ich habe nur genickt, dass es zwei Grenzen gibt. –

Das ist genau das, was Herr Skipka sagt, die 0,2 sind natürlich nicht das δ_{rel}, was häufig als MID herangezogen wird, um dieses δ_{rel} zu definieren. Nein, natürlich ist es δ_{min}; es ist die Irrelevanzschwelle.

Norbert Banik: Ich hatte vorhin ausdrücklich gesagt, dass diese bisher angenommene 0,2 für Sie die Schwelle ist, ab der es sicher relevant ist. Da haben Sie genickt.

Stefan Lange: Nein, da habe ich nicht genickt.

Norbert Banik: Nur aus Verständnisgründen will ich …

Stefan Lange: Nein, da habe ich nicht genickt.

Norbert Banik: Ach so. Es konnte leider nicht protokolliert werden. Das ist schade, weil das sozusagen die Basis wäre, die wir bisher hier am Tisch immer diskutiert haben, zu sagen: Alles drüber betrachten wir als relevant.

Stefan Lange: Nein. Jedenfalls war ich nicht dabei, sagen wir mal so. – Noch mal: Wir reden über die beiden Differenzen, das sicher Irrelevante und das sicher Relevante. Dazwischen gibt es typischerweise eine Grauzone. Das ist das, was Victor gesagt hat. Herr Leverkus hat Victor 1987 angesprochen. Er hat es ja auch schon gesagt und uns vorgeschlagen, dass man sinnvollerweise für die Frage der Relevanz gegenüber dem δ_{min}, also dieser Irrelevanzgrenze, einen verschobenen Test machen sollte. Das ist genau unsere Ansicht.

Guido Skipka: Ich wollte noch einen Aspekt hineinbringen. Im Prinzip ist das nun die richtige Diskussion: Wo legt man jetzt diese Grenze fest? Denn dieser Vorschlag von Kieser / Hauschke hat erst mal einen gewissen Charme, aber man kann das auch anders auffassen. Also: Wenn der Schätzer oberhalb einer gewissen Grenze liegt, ist das nicht gleichbedeutend damit, dass die untere Grenze des Konfidenzintervalls irgendwo bei null liegt. Das ist auch größer.

(Norbert Banik: Kann!)

Herr Dannehl – ich weiß nicht, ob Sie ihn kennen – hat das mal persönlich in Korrespondenz mit uns aufgezeigt. Vielleicht kommen wir am Freitag auch dazu. Ich weiß nicht, ob Sie da sind. Sonst können wir es vielleicht nachher in der Pause mal ansprechen.

Was ich sagen will: Man kann zeigen, dass der Kieser-Hauschke-Ansatz, in dem der Schätzer über der Schwelle sitzt, in etwa damit korrespondiert, dass die untere Grenze des Konfidenzintervalls die Hälfte dieser Schwelle überschreitet. Letztendlich ist es also implizit gar nichts anderes als eine verschobene Hypothese, und man landet sofort wieder bei der eigentlich interessierenden Frage: Wohin soll man die Hypothesengrenze schieben?

Friedhelm Leverkus: Ich bin da gar nicht im Konsens.

Stefan Lange: Im Konsens oder Dissens?

Friedhelm Leverkus: Im Dissens. – Einmal hoffe ich, dass wir bei vielen Bewertungen dieses Problem nicht diskutieren müssen, weil wir Responderraten haben. Aber wenn wir in so einer dummen Situation sind, in der wir so etwas nicht vorliegen haben, halte ich eine formale Entscheidung mit einem statistischen Test auf Relevanzgrenzen, Irrelevanzgrenzen nicht für angebracht, weil diese innerliche Unsicherheit über diese Grenzen so groß ist, dass man da keinen statistischen Test anwenden sollte.

Stefan Lange: Dann würde ich Sie auch bitten, in Ihrer Firma weiterzuverbreiten: Bitte keine Nichtunterlegenheitsstudien mehr! Denn die machen genau das Gleiche.

(Friedhelm Leverkus: Ja!)

Sind Sie meiner Ansicht?

Friedhelm Leverkus: Nein. Es gibt natürlich Dinge, bei denen das ein Konsens ist, dass es so einen Irrelevanzbereich gibt, bei denen es eine Abstimmung darüber gibt. Das mag es irgendwo geben. Deshalb sage ich nie: nie Nichtunterlegenheit. Es mag Situationen geben, in denen das gut begründet ist. Aber Nichtunterlegenheitstests sind wirklich schwer zu interpretieren. Da bin ich auch auf Ihrer Seite.

Stefan Lange: Ja, okay. Dann nehmen wir das erst mal so zur Kenntnis.

Friedhelm Leverkus: Es gibt keinen Nichtunterlegenheitstest mit Cohen's d. Oder haben Sie schon mal so etwas gesehen?

Stefan Lange: Ja, habe ich auch schon gesehen. Ich halte das mit dem Cohen's d, ehrlich gesagt, für ein bisschen zweitrangig, weil sich alle über das Odds Ratio freuen. Das ist nichts anderes als Cohen's d. Das ist ein Maß – ich bin kein Statistiker –, mit dem Biometriker gerne operieren, weil es sehr schöne Eigenschaften hat, und da wird diese Frage typischerweise nicht gestellt.

(Friedhelm Leverkus: Es geht um 0,2 – Entschuldigung!)

Okay. Das können wir in Odds Ratio umrechnen, wenn es Ihnen angenehmer ist.

Friedhelm Leverkus: Das kann man umrechnen, aber es geht ja um die Relevanzgrenze, nicht um Cohen's d, Odds Ratio oder Hedges' g oder so etwas. Das ist klar. Cohen's d hat auch Vorteile, weil er verschiedene Dinge einfach standardisiert. Sie können in einer Meta-Analyse Dinge auch gemeinsam auswertbar machen. Das ist sicherlich ein Vorteil, den ich gar nicht bestreiten will.

Norbert Banik: Vielleicht versuche ich es doch einfach mal mit einer Frage! Das heißt also dann nach Ihrer neuen Einschätzung, dass Sie mit der bisherigen Anwendung des Kriteriums

„statistisch signifikant, größer als standardisierte Differenz 0,2“ auch irrelevante Effekte einbezogen haben,

(Stefan Lange: Nein!)

wenn Sie jetzt plötzlich sagen, das ist gar nicht die Schwelle, ab der es sicher relevant ist?

Stefan Lange: Noch mal: Das ist erst mal kein neues Vorgehen, sondern die Interpretation in der Nomenklatur, die Sie auch in Ihrer Stellungnahme gebracht haben: Es gibt zwei δ, eines, das den sicher irrelevanten Bereich darstellt, und eines, das den sicher relevanten Bereich darstellt. So war das Verständnis bisher. So soll es auch bleiben, dass natürlich gegen das, was den sicher irrelevanten Bereich abgrenzt, getestet wird. Alles andere – da stimme ich zu – könnte auch von der Machbarkeit her problematisch werden. Deswegen: Was man dann sicher hat, ist, dass man das Irrelevante ausgeschlossen hat, möglicherweise in dem Graubereich liegt, aber auch mit einer gewissen Wahrscheinlichkeit im tatsächlich relevanten Bereich – also ganz im Victor’schen Sinne.

Norbert Banik: Und Sie haben kein Problem mit diesen Dingen, die wir auch in der Stellungnahme mal beispielhaft gebracht haben, die im Cochrane Handbook und woanders auch stehen, dass man leicht auch Effektstärken in Number Needed to Treat, in Odds Ratios usw. übersetzen kann und es durchaus nicht ausgemacht ist, dass diese Schwelle 0,2 sein muss, unabhängig von dem Kontext, sondern dass zu gleichen Numbern Needed to Treat auch standardisierte Effektmaße von 0,1 passen können, wie sie zu 0,2 passen, entsprechend abhängig von der Situation, Eventhäufigkeit oder Änderungsverteilung des jeweiligen Parameters.

Stefan Lange: Klar. Das kann schon sein. Das gilt auch für das Odds Ratio oder für das relative Risiko – ein schönes Maß, das wir in klinischen Studien immer gerne nehmen.

Guido Skipka: Es ist vollkommen klar, dass, wenn ich eine NNT festsetze, komme ich, je nachdem wo die Verteilungen liegen, zu unterschiedlichen standardisierten Mittelwertdifferenzen. Ich kann aber auch die SMD festsetzen und zu unterschiedlichen NNTs oder Odds Ratio kommen. Das liegt einfach daran, dass die Distanzmaße meist noch von einer dritten Größe abhängen. Da sehe ich, ehrlich gesagt, nicht so ganz das Problem. Wie gesagt, diese 0,2, bezogen auf die SMD, wollen wir in Ermangelung irgendwelcher ankerbasierten Informationen anwenden. Das ist meiner Meinung nach auch nicht durch den Wechsel zu einem anderen Distanzmaß wie NNT zu lösen.

Norbert Banik: Wie auch immer, das ist uns klar, da stimmen wir auch überein. Aber es geht im Grunde genommen darum, wenn man es doch über diese Schwelle machen will, zu einer Festlegung zu kommen, welche es sein soll. Wenn es so ist, dass 0,2 erstens nicht universell ist, Sie jetzt aber sagen, dass 0,2 für Sie die Grenze der Irrelevanz ist, also dieses δ_{irr}, dann heißt das aber doch, dass Sie wirklich davon ausgehen, dass Effekte, die kleiner als 0,2 sind,

irrelevant sind. Deshalb habe ich noch mal dieses Beispiel gebracht. Dem kann in der praktischen Anwendung einfach nicht so sein. Das hat auch weder Cohen noch irgendjemand anders je gesagt. Damit machen Sie wirklich einen großen zusätzlichen Schritt. In der Vergangenheit hatte ich noch nicht mal vermutet, dass Sie den Schritt machen, weil ich immer gedacht habe: Aha, das ist einfach die fehlende Unterscheidung zwischen Irrelevanzgrenze und Relevanzgrenze. Aber wenn Sie jetzt sozusagen noch weiter hoch gehen und sagen, es ist sogar so, dass alle Effekte unter 0,2 irrelevant sind, bin ich sehr überrascht und ich weiß nicht, ob Sie das unterschreiben würden, wenn man es auf diesen Punkt bringt.

Stefan Lange: Noch mal: Wir machen ein hierarchisches Vorgehen. Gibt es eine MID? Gibt es Responderanalysen? Dann nehmen wir Responderanalysen. Gibt es eine MID, werden wir versuchen, auf eine solche Irrelevanz- oder Relevanzschwelle … Ich habe gerade noch mal unsere Nomenklatur nachgelesen, da kann man noch ein bisschen nacharbeiten. Es ist auch denkbar: 0,2 ist nicht in Stein gemeißelt. Da stimme ich Ihnen völlig zu. Es kann auch sein, dass in Abhängigkeit von der MID ein anderes δ_{irr} definiert und festgelegt werden kann. Aber irgendein δ_{irr} wird es geben – das haben Sie auch gerade gesagt –, wobei Sie in Ihrer Stellungnahme geschrieben haben, Sie gehen davon aus, dass das δ_{irr} immer gleich null ist. – Nur, das macht keinen Sinn.

(Norbert Banik: Mit einer bestimmten Begründung des Kontextes!)

Das sind patientenrelevante Endpunkte. Aber wir reden ja nur über patientenrelevante Endpunkte.

(Norbert Banik: Nein, wir reden in dem Kontext Kosten-Nutzen-Bewertung, den Sie nicht hören wollen, obwohl er da ist! Aber gut!)

Nein, wir reden nicht im Kontext Kosten-Nutzen-Bewertung. Wir reden im Kontext Nutzenbewertung und Relevanzbewertung. Das hat zunächst mal nichts mit Kosten-Nutzen-Bewertung zu tun.

Norbert Banik: Wieso hat es damit nichts zu tun? Wenn Sie doch nur die Dinge zur Kosten-Nutzen-Bewertung zulassen, die bereits einen Nutzennachweis haben, und mit diesem Kriterium der Nutzennachweis stehen oder fallen kann, dann hat es doch etwas damit zu tun. Das kann man doch nicht negieren.

Stefan Lange: Ja, aber wir reden heute nicht über die Kosten-Nutzen-Bewertung.

(Norbert Banik: Ach so, das ist die Maulkorbvariante!)

Herr Banik, ich glaube …

Norbert Banik: Wir müssen darüber reden, weil es Konsequenzen für Werturteile hat. Das versuchen wir ja zu erklären.

Stefan Lange: Da würde ich Sie bitten, doch immer noch freundlich und höflich zu bleiben, weil – das hatte ich am Anfang gesagt – wir hier nicht über die grundsätzlichen Dinge sprechen können. Das sprengt den Rahmen. Für uns war die Frage: Wie können wir im Rahmen einer Nutzenbewertung eine Relevanzbewertung machen? Da wünschten wir uns gerne den Austausch, Herr Banik.

Norbert Banik: Es ist richtig, und ich möchte auch gar nicht übers Ziel hinausschießen. Nur, wir haben nicht die Möglichkeit, dieses Thema generell zu diskutieren, weil es in den bisherigen Methodenpapieren nicht verankert war. Da hätten wir es generell diskutiert. Deswegen diskutieren wir es jetzt peu à peu bei jeder Nutzenbewertung. Es nutzt leider für die Festlegung des Kriteriums nichts, dass Ihnen dieser Rahmen gegeben ist, dass sie jetzt den Auftrag haben, Nutzenbewertungen zu machen, die bei nachgewiesenem Nutzen in eine Kosten-Nutzen-Bewertung überführt werden.

Da kommt das Werturteil der Versichertengemeinschaft zum Tragen zu sagen: Zahlen wir dafür oder zahlen wir nicht? Wenn Sie vorher sagen, alle Effekte unter 0,2 sind irrelevant und ich will sie dir, egal, ob sie etwas kosten oder nicht, überhaupt nicht vorlegen, finde ich das bedenklich. Denn damit wird diesmal wirklich – vielleicht erstmalig in unseren jahrelangen Diskussionen – explizit ein patientenrelevant sein könnender Effekt vorenthalten. Diese Konsequenz muss man sich bewusst machen, und man muss sich auch Größenordnungen bewusst machen. Deshalb habe ich Sie gefragt: Können Sie unterschreiben: Alles unter 0,2 ist für Sie generell und immer in den Verfahren, in denen es bisher angewandt wurde, irrelevant? Ich habe noch nicht gehört, dass Sie Ja gesagt haben. Aber Sie müssten Ja sagen, sonst würde diese Schwelle nicht diese Rolle einnehmen können.

Stefan Lange: Nein – ich sage es zum dritten oder vierten Mal –, wir haben die Hierarchie. Das wissen Sie ja auch; Sie kennen unsere Berichte.

(Norbert Banik: Wir sind auf der Ebene der Hierarchie, wo es zur Anwendung kommt!)

Im Bereich der Antidepressiva haben wir Responderanalysen gehabt, die wir auch für die Nutzenbewertung herangezogen haben, und sind auch zu entsprechenden Nutzenbelegen gekommen. Wo wir sie nicht haben und es keine MID gibt, müssen wir irgendein Kriterium heranziehen; es hilft ja nichts. Wir drehen uns in der Diskussion sozusagen im Kreis. Sie weigern sich standhaft, zu sagen: Dann nehmen wir nicht 0,2. Wir könnten von mir aus auch anhand einer konkreten Skala sprechen. Sollen wir mal – welches Anwendungsgebiet schlagen Sie vor? – über Hamilton sprechen?

(Norbert Banik: Schlechtes Beispiel!)

Schlechtes Beispiel. Sagen Sie ein besseres Beispiel! Nehmen wir doch mal eine Skala!

Norbert Banik: Häufigkeit nächtlichen Aufwachens.

Stefan Lange: Häufigkeit nächtlichen Aufwachens, das ist aber eine nicht so furchtbar komplexe Skala. Wir reden ja gerade über komplexe Skalen im Bereich Lebensqualität – das wäre schon hilfreich – oder Symptommessungen, aber wo es ein bisschen komplexer ist. Bei Alzheimer wäre das ADAS-cog oder so etwas.

Norbert Banik: Jetzt reden wir ja konkret über COPD, und da fällt mir keine komplexe Skala ein, für die das nicht ohnehin existiert. Wir haben oft St. George erwähnt.

Stefan Lange: Gut. Dann nehmen wir doch St. George! Was ist denn bei St. George irrelevant? Wie viele Punkte? Nicht in Standardabweichung, sondern wie viele Punkte im St. George sind irrelevant? Welches könnte dort das δ_{irr} sein?

Norbert Banik: Ich bin nicht der Erfinder der Skala St. George, der bekanntlich Paul Jones ist und meines Wissens publiziert hat, dass es zwischen drei und vier Punkten liegt. Unterhalb dessen ist es also irrelevant.

Stefan Lange: Okay. Das heißt, da würden Sie jetzt sagen, das könnte man, wenn man das wüsste, als einen Anker für eine verschobene Nullhypothese verwenden.

Norbert Banik: Das brauchen Sie in dem Falle nicht; da haben Sie ja die Relevanzschwelle. Sie sind ja in der Hierarchie jetzt eine Ebene drüber.

Stefan Lange: Aber wenn wir keine Responderanalysen für St. George haben, was machen wir dann?

Norbert Banik: Dann messen wir den beobachteten Effekt relativ zu dieser vielleicht anerkannten Schwelle. Ich möchte mich da nicht zu weit rausbeugen.

Stefan Lange: Nehmen wir mal an, sie sei drei Punkte! Dann können wir sagen: Wenn die untere Grenze vom Konfidenzintervall diese drei nicht ausschließen kann, dann müssen wir davon ausgehen, dass es nicht nicht sicher irrelevant ist. Also umgekehrt können wir nur sichergehen, wenn die untere Grenze des 95%-igen Konfidenzintervalls oberhalb dieser drei Punkte ist. Stimmt das? Habe ich Sie jetzt richtig verstanden?

Norbert Banik: Ja, wenn das Kriterium in dem Falle existiert, können Sie das so machen – wie bisher auch.

Friedhelm Leverkus: Es existiert das Kriterium, dass drei Punkte möglicherweise

(Norbert Banik: Das ist intraindividuell!)

klinisch relevant sind. Die Frage, die sich noch stellt …

Stefan Lange: Ich frage nach dem δ_{irr}, nicht nach dem δ_{rel}.

Norbert Banik: Da habe ich Sie missverstanden, das δ_{irr} kenne ich nicht.

Stefan Lange: Aber darauf müssen wir doch hinaus.

Norbert Banik: Wir haben jetzt auf individuellem Niveau gesprochen.

Stefan Lange: Nein, auf Gruppenebene. Was ist das δ_{irr}?

Norbert Banik: Das kann ich nicht sagen.

Stefan Lange: Aber das hilft uns nichts. Wir sind jetzt im Szenario: Wir haben keine Responderanalysen. Wir haben nur Mittelwertunterschiede. Nun müssen wir sagen, ob das relevant ist oder nicht. Wir brauchen ein δ_{irr}. Sagen Sie es mir!

Norbert Banik: Ich kann nur sagen: Für St. George kenne ich das δ_{irr} nicht.

Stefan Lange: Gut. Das ist sehr hilfreich. Danke.

Friedhelm Leverkus: Ich kenne das auch nicht. Ich kenne nur die Diskussion, in der es drum geht, ob das nun auf der Gruppenebene oder Einzelebene ist.

Stefan Lange: Noch mal: Wir sind in der Situation, und da hilft es uns nicht, wenn man sagt: Das müssen wir erst mal in einen Kontext stellen und da noch mal bewerten. Wir brauchen das δ_{irr}.

Friedhelm Leverkus: Herr Lange, wenn wir mal in Respondern denken.

Stefan Lange: Nein, Responderanalysen haben wir nicht.

Friedhelm Leverkus: Aber wenn Sie eine Responderanalyse hätten, würden Sie die Entscheidung treffen und sagen: Da sind 20 %, da sind 30 %. Das, was in der Skala passiert, beeinflusst ja die Responderanalysen. Der Mittelwert beeinflusst im Prinzip die Responder. Wäre es denn nicht eine Möglichkeit, dass Sie dann sagen: Wenn ich aus irgendeinem Grund nicht an eine Responderanalyse herankommen kann, versuche ich, das über ein Normalverteilungsmodell abzuschätzen? Ich glaube, Herr Skipka, das haben Sie sogar mal gemacht, dass man versucht zu sagen: Okay, ich habe jetzt wahrscheinlich den Anteil Responder und den Anteil Responder.

Guido Skipka: Nur kurz als Einschub: Wir haben das gemacht, aber ich meine nicht, um daraus eine Nutzenaussage zu treffen, sondern im Rahmen einer Diskussion. Ich bin aber nicht ganz sicher. So etwas wäre eine denkbare Möglichkeit.

Stefan Lange: Dann würde ich mal nachfragen, Herr Leverkus: Wie konstruieren Sie dann aus so etwas statistische Hypothesen und Tests?

Friedhelm Leverkus: Das würde ich hier nicht machen.

Stefan Lange: Aber es bleibt doch nichts übrig. Wir sagen ja: Wir brauchen zumindest eine Signifikanzaussage.

Friedhelm Leverkus: Man könnte es über LCA – Latent Class Analysis – machen, dass Sie im Prinzip sagen, im Hintergrund haben Sie eine binäre Variable, die letztlich so eine stetige Variable erzeugt.

Stefan Lange: Nur, da steckt man wahrscheinlich starke Annahmen rein.

Friedhelm Leverkus: Das müsste man mal sehen.

Norbert Banik: Ich würde die Frage einfach zurückgeben. Sagen wir doch mal, wir haben uns jetzt hier mehr oder weniger geeinigt, wir alle kennen diese Irrelevanzschwelle für St. George's Questionnaire für COPD nicht. Dann würde es automatisch so sein, dass dieses patientenrelevante Maß in die dritte Hierarchieebene fällt, nämlich die, wo man jetzt irgendetwas mit einer standardisierten Mittelwertdifferenz bewerten müsste. Dann ist es im Umkehrschluss einfach so, dass Sie sagen: Aha, alles unter 0,2 standardisierte Differenz im St. George ist nach Ihrer Definition irrelevant.

Stefan Lange: Ja, das könnte passieren.

Norbert Banik: Das muss sicher so sein, sonst dürfen Sie diese Schwelle nicht nehmen.

Stefan Lange: Wenn wir kein δ_{irr} haben, ist das so.

Norbert Banik: Dann ist es als „unter 0,2" immer irrelevant. Das würde ich gerne aufgreifen und mal mit Paul Jones diskutieren. Das ist vielleicht auch eine Möglichkeit für Sie, aber ich würde es machen.

Stefan Lange: Nur, um das zu präzisieren – das hatte ich vorhin auch gesagt, möglicherweise ist das im Berichtsplan nicht so herausgekommen: Die 0,2 wollen wir nur dann verwenden, wenn es keine MID gibt. Wenn es eine MID gibt, so wie ich Sie verstanden habe, kann durchaus auch ein anderes δ_{irr} als die 0,2 herauskommen, nämlich auf Basis der Skala. 0,2 ist ein guter Anhalt. Man kann es sich auch ausrechnen. Ich weiß nicht, wie die Standardabweichung da typischerweise von Veränderungen auf der Skala ist. Aber – das kann ich jetzt nicht auswendig sagen – da bestehen auch noch andere Möglichkeiten, als sich an dem Cohen's d von 0,2 festzuhalten.

Norbert Banik: Wir haben aber gerade gesagt, auf Gruppenebene existiert diese MID nicht. Das ist immer ein bisschen gefährlich, in Abkürzungen zu sprechen, weil jeder etwas anderes darunter versteht. Aber wir haben gerade gesagt, sie existiert nicht. Damit kommen wir also nicht raus. Dass man das eine in das andere umrechnen kann, wissen wir. Das ist natürlich

auch eine interessante Illustration, die wir schon angeregt haben, zu sagen: Man kann es doch umrechnen und einfach gucken, was es bedeutet. Deshalb frage ich immer wieder, warum es gerade die 0,2 sein soll, und drunter ist für Herrn Privatdozenten Lange wirklich alles irrelevant.

Stefan Lange: Ich weiß auch nicht, Herr Banik, ich glaube, dann hören wir jetzt besser auf …

Norbert Banik: Doch, als Irrelevanzschwelle muss es doch diese Interpretation haben.

Stefan Lange: Nein, Herr Banik, wir versuchen doch hier, gut miteinander zu reden, aber dann müssen Sie doch nicht so komische Bemerkungen machen. Das ist einfach völlig überflüssig: für den Herrn Privatdozenten Dr. Lange. Ich sage ja auch nicht: Für den Herrn Dr. Banik ist irgendetwas …

(Norbert Banik: Das war nicht ungut gemeint!)

Stefan Lange: Das war gut gemeint, okay. – Ich hatte Ihnen schon zum zehnten Mal gesagt, das ist nicht immer 0,2. Nur, wenn wir nichts anderes haben, dann bleibt uns nichts anderes übrig, als etwas festzulegen. Dann nehmen wir die 0,2 – für dieses Projekt, im Berichtsplan beschrieben. Und wenn wir etwas anderes haben, nehmen wir etwas anderes. Ich kann es nicht tausendmal sagen. Deswegen können Sie ständig wiederholen, ich würde immer sagen 0,2. Ich sage es einfach nicht.

Guido Skipka: Vielleicht noch eine Bemerkung – vielleicht wird das auch zwischenzeitlich vergessen –, insbesondere wenn Sie die Konsequenzen ansprechen: Wir sehen – das haben wir unter TOP 2a) diskutiert – aus unserer Sicht bei diesen Skalen die Notwendigkeit, eine Relevanzbewertung durchzuführen, um zu einer Nutzenaussage zu kommen. Wenn es nun mal keine Informationen gibt und wir auch nicht unsere willkürliche Grenze 0,2 nehmen sollen, wäre die Alternative, in keinem Fall eine Nutzenaussage zu treffen. Das ist doch eigentlich noch viel konservativer als das, was wir vorschlagen, oder? Es könnte doch mal sein, dass dieses Kriterium auch erfüllt wird; das wäre doch wunderbar für alle.

Friedhelm Leverkus: Das ist besser, als den Nutzen abzusprechen; da gebe ich dir vollkommen recht. Aber aufgrund der Unsicherheit, die wir haben – ich kann mich nur noch mal wiederholen –, erscheint mir die Herangehensweise, Nutzen aufgrund von einem statistischen Test auszuschließen, auch ein bisschen zu konservativ. Das ist der Punkt, um den sich jetzt alles dreht.

Stefan Lange: Gut. Stellen wir also erst mal fest: Im Grundsatz ja, die Notwendigkeit der Relevanzbewertung wird gesehen. Die Schwierigkeit ist nur die Frage nach dem geeigneten Verfahren. So eine richtige Alternative habe ich für eine Entscheidung, die daraus resultieren soll, noch nicht gehört. Das ist unser wichtiger Punkt: Wir müssen eine Entscheidung fällen.

Da hilft uns die Beschreibung nicht. Die Beschreibung ist noch keine Entscheidung, sondern es braucht dann eine Entscheidungsregel. Die Entscheidungsregel habe ich noch nicht gehört.

(Friedhelm Leverkus: Irgendetwas anderes!)

Ja, deshalb fragen wir ja die ganze Zeit.

Friedhelm Leverkus: Die Entscheidungsregel von uns ist eher, dass wir sagen: Okay, das ist ein Werturteil, und das Werturteil gehört eigentlich nicht in die Nutzenbewertung, sondern …

Stefan Lange: Also, alles, was signifikant ist, ist gut, weil das Werturteil im Rahmen der Nutzenbewertung hier nicht getroffen werden kann, sondern das würden Sie bei andern Personen sehen. Okay.

Friedhelm Leverkus: Nein, von Ihnen kann auch eine Kosten-Nutzen-Bewertung durchgeführt werden; das kann ja auch passieren.

Stefan Lange: Aber nicht im Rahmen der Nutzenbewertung, so habe ich Sie gerade verstanden. Dann ist es also im Grunde genommen das, was Herr Banik auch in seiner Stellungnahme geschrieben hat: Für patientenrelevante Endpunkte gibt es kein δ_{irr} beziehungweise dieses δ_{irr} ist immer gleich null. – Das hilft uns hier jetzt nicht so viel, aber dann ist das Ihr Vorschlag oder Ihre Antwort auf unsere Frage, wie man es alternativ machen könnte.

Harald Kögler: Zurück zur Semantik. Ich verstehe Sie richtig, wenn δ_{irr} von 0,2 erfüllt ist, sprechen Sie von einem Beleg für den Zusatznutzen. Nun unterscheiden Sie ja zwischen Beleg für einen Zusatznutzen und Hinweisen auf einen Zusatznutzen. Hinweis wäre für mich eine schwächere Aussage als ein Beleg. Wenn 0,2 als ein Beleg angesehen wird, wo sehen Sie dann die Grenze für einen Hinweis auf einen Zusatznutzen?

Guido Skipka: Diese Abschwächung von einem Beleg für einen Nutzen oder Zusatznutzen auf einen Hinweis zu einem solchen machen wir nicht an der Quantität der Effektstärke fest, sondern an der Qualität der Ergebnisse. Also, wenn wir jetzt sehen, die Studien haben ein gewisses Verzerrungspotenzial, dann schwächen wir deswegen gegebenenfalls auf einen Hinweis ab.

Stefan Lange: Gut. Wir sind nicht viel, vielleicht ein Stückchen weitergekommen. Trotzdem vielen Dank für die dennoch interessante Diskussion. – Wir kommen noch zu:

4.3.8 Tagesordnungspunkt 3: Verschiedenes

Peter Kardos: Ich hätte eine Frage. Im Berichtsplan steht, dass die Endpunkte, die sich auf weniger als 70 % der Patienten beziehen, wegen Drop-outs ausgeschlossen werden sollten. Ist das so oder galt das so? Denn dagegen spricht natürlich die Natur dieser Langzeitstudien. Je länger eine Studie ist, umso wertvoller ist das für die klinische Beurteilung dieses Arzneimittels bei einer Erkrankung, die ja eine chronisch-progrediente Erkrankung ist. Die Endpunkte würden dann ausfallen.

Guido Skipka: Das Ganze beruht wahrscheinlich auf einem Missverständnis oder einer Missdeutung der Begriffe. Wir meinen mit Nichtberücksichtigungsrate den Anteil der Patienten, die überhaupt nicht in die Auswertung einbezogen werden. Es ist unbestritten, dass gerade bei Langzeitstudien Patienten ausscheiden. Es kann ja auch keiner gezwungen werden, in der Studie zu bleiben. Das heißt noch nicht, dass man diese Patienten aus der Auswertung ausschließen muss. Zum Teil sind zumindest über einen gewissen Zeitraum Werte da. Man kann sich dann Gedanken über Ersetzungsstrategien machen. Dazu gibt es die ganze ITT-Diskussion. Aber Patienten ganz herauszulassen, da sehen wir ein Problem, weil das einfach Verzerrungspotenzial erzeugt. Wir haben eine Grenze festgesetzt, ab der wir sagen: Das ist einfach so viel mit den Ergebnissen, die halten wir nicht mehr für aussagekräftig. Noch mal: Es geht nicht um Therapieabbrecher. Das können von mir aus aus statistischer Sicht irgendwann 100 % sein. Solange die weiter beobachtet sind, habe ich da erst mal kein Problem. Es mag vielleicht klinisch zur Beantwortung der entscheidenden Fragen ein Problem sein. Aber es geht erst mal nicht um Ausfälle aus der Studie; es geht darum: Wer ist letztendlich in die Auswertung einbezogen worden.

Peter Kardos: Wenn ich die Diskussion als blutiger Laie in der Statistik beobachte, stelle ich doch fest, dass auch CONSORT mittlerweile von dieser strikten ITT-Geschichte abgekommen ist. Sie haben gesagt, Sie müssen nur wissen, wer dann in die Auswertung mit hineingenommen worden ist. Als diese Langzeitstudien geplant worden sind, die wir heute als Meilenstein ansehen, war diese Geschichte mit der weiteren Beobachtung der Drop-outs eigentlich noch nicht so angenommen und nicht so klar.

Ich habe mir als Kliniker immer schon Gedanken darüber gemacht: Wieso soll ich einen Patienten, der zwei Wochen lang mit einem Medikament behandelt worden ist und dann 24 oder 35 Monate nicht, im Rahmen der ursprünglichen Zuordnung in die ITT-Auswertung nehmen? Ich habe den Eindruck, dass das mittlerweile sehr umstritten ist und wirkliche Vorreiter dieser ITT-Empfehlung, wie zum Beispiel die kanadische Gruppe um Suissa und Aaron, inzwischen Artikel schreiben, in denen steht: Wenn der Patient nur zwei Wochen behandelt worden ist und nach drei Jahren ausgewertet wird, dann ist das falsch; so eine Studie kann man nicht verwenden. – Das hat mich übrigens maßlos überrascht.

Ich habe den Eindruck, dass diese nachträglich aufgestellten Kriterien – so muss ich das sagen – vielleicht nicht angewendet werden sollen. Ich kann nicht abwägen und nicht einschätzen,

wie viele Endpunkte von der Bewertung ausgeschlossen werden. Und die 20 %- oder die 30 %-Grenze – das habe ich auch geschrieben – bezieht sich natürlich nicht auf solche extrem lang laufenden Studien, von denen wir jetzt sprechen. Das sind kürzere Studien.

Beate Wieseler: Ich glaube tatsächlich, dass es primär um ein Missverständnis geht. Es geht nicht mehr primär darum, dass ich die Patienten tatsächlich alle bis zum Schluss weiter beobachte, sondern darum, dass ich diese Patienten in irgendeiner geeigneten Form in der Auswertung berücksichtige.

(Peter Kardos: Fortführung des letzten bekannten Wertes!)

Zum Beispiel, wenn das ein adäquates Verfahren für diesen Endpunkt ist, wäre das eine Möglichkeit.

Stefan Lange: Ist Ihre Frage so weit beantwortet?

(Peter Kardos: Ja!)

Ich glaube auch, dass es da tatsächlich ein Missverständnis gab.

Norbert Banik: Ich wollte noch zwei Fragen zu Sonstigem stellen. Ihnen ist da offenbar alles klar gewesen. Aber wir würden doch gern wissen, wie es mit dem Umfang der Bewertung steht. Denn unseres Erachtens – das hatten wir auch dargelegt – war die Frage einzeln oder in Kombination für Tiotropium relativ unklar, für die Vergleichsmedikation vielleicht etwas klarer, und wir wollten gern wissen, um das abschätzen zu können, wie Sie den Umfang tatsächlich sehen.

Beate Wieseler: Wir werden alle Interventionen, die gemäß der Zulassung in Deutschland in den Studien eingesetzt werden, auch bewerten. Es gibt ja auch für Tiotropium erst mal primär in der Zulassung keine Einschränkung bezüglich der Kombinationen, sodass wir das bewerten würden.

Stefan Lange: Frage beantwortet?

Norbert Banik: Ja. – Die zweite kleine Nachfrage, von Ihnen nicht zum Gegenstand erhoben, deshalb vielleicht klar, wäre noch die Frage nach der minimalen Studiendauer. Wir hatten vorgeschlagen, dass bei schweren Graden der COPD unseres Erachtens auch bei kürzeren Studiendauern relevante Effekte abbildbar sind, und fragen, ob es bei der minimalen sechsmonatigen Studiendauer – übersetzt: 24 Wochen – bleibt oder ob Sie aufgrund dieser Darlegungen auch kürzere Studiendauern mit einbeziehen wollen.

Beate Wieseler: Wir sehen im Moment keinen Grund, kürzere Studiendauern zu berücksichtigen. Es werden eigentlich generell – auch für die Symptomatik – sechs Monate veranschlagt, um da zu einer sicheren Aussage zu kommen. Es gibt auch einzelne Studien, in

denen transiente Effekte, zum Beispiel für Salmeterol, in den ersten drei Monaten gezeigt werden. Wir denken, dass eine Studiendauer von sechs Monaten innerhalb dieser Indikation berechtigt ist – auch für alle Endpunkte. Denn wir sehen durchaus in den ersten Monaten eine Entwicklung in den Endpunkten, und es liegt uns nicht daran, in hohem Ausmaß falsch negative Studien in der Bewertung zu haben, weil die Effekte bei COPD im Vergleich zu Asthma wiederum später sichtbar werden.

Norbert Banik: Nur als Nachfrage: Das wollen Sie also auch bei COPD-Patienten mit schwerem und sehr schwerem Schweregrad beibehalten, bei denen zum Beispiel Exazerbationshäufigkeiten bereits in kürzeren Studien nachgewiesen sind, da die Wahrscheinlichkeit, dass der Effekt transient ist, in diesem präfinalen Stadium per se sehr unwahrscheinlich ist, würde ich mal sagen. Das wollen Sie wirklich in allen Situationen durchziehen, dass es nur Studien mit einer minimalen Dauer von sechs Monaten gibt?

Beate Wieseler: Ja, aus heutiger Sicht.

Peter Kardos: Wenn es Studien mit Stadium-IV-Patienten gäbe, die ganz selten sind, wenn es sie überhaupt gibt, ich könnte keine zitieren, dann mag das schon berechtigt sein, Vier- und Sechs-Wochen-Effekte zu berücksichtigen. Aber wenn Sie nur an die Exazerbationen etwa mit den saisonalen Effekten denken, hätte ich schon Probleme damit. Sie müssen ja bedenken, dass die IVer-Patienten, bei denen es relevant wäre, grundsätzlich von den Studien ausgeschlossen worden sind. Dieses Vorgehen ist auch schon kritisiert worden. Eigentlich haben wir keine Evidenz für die Behandlung von GOLD-IV-COPD-Patienten, von sauerstoffpflichtigen COPD-Patienten.

Thomas Glaab: Frau Wieseler, eine Frage noch zur Intervention bei Tiotropium. Sie sprachen davon, dass Sie dort alle Interventionen für möglich halten. Das heißt: Verstehe ich das richtig, dass Sie auch Tiotropium versus Fixkombis ICS/LABA in diesem Auftrag sehen? Das wäre mir neu.

Beate Wieseler: Wie gesagt, der Auftrag ist, Tiotropium mit anderen verfügbaren, in Deutschland zugelassenen Therapieoptionen zu vergleichen. Die Fixkombinationen sind für COPD zugelassen. Meine Frage an Sie – Sie sprechen wahrscheinlich den Vergleich Tiotropium-Mono versus Fixkombination Kortikosteroid / langwirksames Beta-2-Sympathomimetikum an:

(Thomas Glaab: Genau!)

Wie beurteilen Sie diese Fragestellung aus klinischer Sicht? Ist das ein relevanter Vergleich, unabhängig davon, dass beide Präparate in dieser Form in Deutschland zugelassen sind?

Thomas Glaab: Ja, gut, ich sehe schon einen deutlichen Unterschied – auch in der Zulassung. Die Fixkombination ICS/LABA oder überhaupt inhalative Steroide, gerade aber

die Fixkombis sind erst ab Stadium III und IV zugelassen plus wiederholte Exazerbationen, also eine sehr eingeschränkte Indikation. Da wundere ich mich schon, dass man eine Monotherapie mit einer Fixkombi vergleichen möchte, weil mein Verständnis eigentlich gewesen wäre, hier Tiotropium, langwirksame Beta-2-Mimetika – was gäbe es noch? –, Theophyllin, was auch immer, zu kombinieren. Deswegen auch die Nachfrage, weil ich Sie bitte, das noch mal zu klären.

Peter Kardos: Sie haben gesagt: klinische Sicht. Aus klinischer Sicht ist das schon so, dass die beiden Präparate gegeneinander laufen auf dem Markt – keine Frage. Die Kombination ist aus wissenschaftlicher Sicht, wie Sie, Herr Glaab, das gerade richtig gesagt haben, völlig irrelevant und die Indikation ist eine ganz andere. Die ICS haben nur eine eingeschränkte Indikation. Das tägliche Leben ist nicht so. – Ich möchte mich verabschieden; ich muss leider in drei Minuten gehen. Ich bedanke mich für die Teilnahme.

Stefan Lange: Wir bedanken uns, dass Sie gekommen sind und uns hier unterstützen.

Thomas Glaab: Zu Therapieindikationen: Was ich verstehe, ist die Begleitmedikation inhalative Steroide zu Tiotropium-Mono versus langwirksame Beta-2-Mimetika plus Begleitmedikation ICS. Aber dass man hier plötzlich ein Monopräparat mit Fixkombi vergleicht, das möchte ich doch hinterfragen.

Harald Kögler: Es ist mir auch nicht erinnerlich, dass Sie beispielsweise bei Antidepressiva Kombinationspräparate mit Einzelwirkstoffen verglichen hätten. Ich finde diesen Ansatz sehr ungewöhnlich. Denn eine Kombination mit zwei unterschiedlichen Wirkprinzipien hat einfach eine andere Ausgangssituation als ein Einzelwirkstoff.

Beate Wieseler: Wie gesagt, wir werden das im Rahmen der gültigen Zulassung betrachten. Wenn Sie zum Beispiel da eine Studienpopulation hätten, die für ICS/LABA nicht relevant ist, würden wir auch diese Studie nicht betrachten. Wir würden uns also für jede der Interventionen anschauen: Ist diese Studie im Rahmen der gültigen Zulassung? Ganz unabhängig davon gebe ich Ihnen recht, dass ich auch aus wissenschaftlicher Sicht eine Studie Tiotropium versus langwirksame Betamimetika Add-on und zu Kortikosteroiden für zielführender halten würde. Aber wenn diese Studie da ist, die Sie skizziert haben, würden wir uns anschauen, wie sich das zur Zulassung verhält, die diese Präparate in dieser Population haben.

Thomas Glaab: Ich glaube, da nähern wir uns vom Verständnis an. Die Studien, von denen wir sprechen, sind ab Stadium II bis IV. Es sind sehr viele Stadium-II-Patienten drin, wo wir sagen, aufgrund von Leitlinienempfehlungen und entsprechend der Zulassungssituation ist es so, dass wir hier wirklich erst ab fortgeschrittener COPD und dann noch wiederholten Exazerbationen die Indikation für inhalative Steroide wirklich haben, on top zu langwirksamen Bronchodilatatoren.

Cordula Hagedorn: Ich teile das im Prinzip und möchte nur zu bedenken geben, dass es für die in Deutschland vorhandenen ICS/LABA-Kombinationen auch Unterschiede in der Zulassung gibt, obwohl die meisten ab 50 % FEV_1 zugelassen sind, aber auch eine vorhanden ist, die ab 60 % bereits zugelassen ist. Das wäre entsprechend wichtig zu bedenken.

Norbert Banik: Meine Frage war primär nach dem Verständnis: Wie weit wird die gesamte Bewertung überhaupt aufgefächert? Das hat auch mit der Recherche und mit dem Einstieg zu tun. Was dann später womit verglichen wird, ist eine andere Frage. Es ist guter Standard hier im Haus, dass Sie das sowieso nach der deutschen Indikation und Zulassung machen. Deswegen wollte ich die Frage stellen: Wie weit ist es? Alles andere wird entsprechend aufgearbeitet.

Stefan Lange: Ist Ihre Frage damit beantwortet?

Norbert Banik: Ja.

Stefan Lange: Prima. Damit sind wir durch, wenn nicht noch andere dringende Fragen sind. Das passt auch, weil wir uns einen Rahmen bis ca. 16 Uhr gesetzt hatten. Ich bedanke mich für die sehr interessante und engagierte Diskussion und wünsche Ihnen einen guten Nachhauseweg.

Anhang: Dokumentation der Stellungnahmen

Inhaltsverzeichnis

A 1 Stellungnahmen von Organisationen, Institutionen und Firmen

A 1.1 Boehringer Ingelheim Pharma GmbH & Co. KG

Autoren:
Glaab, Thomas, PD Dr.
Kögler, Harald, PD Dr.
Leimer, Inge, Dr.

Adresse:
Boehringer Ingelheim Pharma GmbH & Co. KG
Binger Straße 173
55216 Ingelheim

Stellungnahme der Firma Boehringer Ingelheim Pharma GmbH & Co. KG

(im Folgenden „Boehringer Ingelheim“) zum vorläufigen Berichtsplan der Bewertung A05-18 „Tiotropiumbromid bei COPD“

Einleitung

Das IQWiG hat am 19.04.2010 den vorläufigen Berichtsplan zum Auftrag A05-18 „Tiotropiumbromid bei COPD“ veröffentlicht.

Boehringer Ingelheim kommentiert hiermit den vorläufigen Berichtsplan.

Wir begrüßen ausdrücklich die Einbindung von Patientenorganisationen in den vorliegenden IQWiG-Prozess. Angesichts der grundlegenden Unterschiede zwischen der COPD und dem Asthma bronchiale ist jedoch nicht nachvollziehbar, warum während der Erstellung des Berichtsplans als Patientenvertreter Mitglieder des Deutschen Allergie- und Asthmabundes e.V. einbezogen wurden, auf dessen Internetseiten unter http://www.daab.de/index.php keinerlei Hinweis darauf zu finden ist, dass der DAAB eine Interessenvertretung von COPD-Patienten darstellt. Patientenverbände, die sich explizit den Belangen von COPD-Patienten verschrieben haben, wie beispielsweise die Selbsthilfegruppe Lungenemphysem COPD Deutschland (http://lungenemphysem-copd.de/), oder die COPD Selbsthilfe e.V. (http://copd-selbsthilfe.de) wurden hingegen nicht konsultiert.

1 Präzisierung der Zielgröße “gesundheitsbezogene Lebensqualität”

Im vorläufigen Berichtsplan wird als eine der Zielgrößen “gesundheitsbezogene Lebensqualität” genannt. Dieser patientenberichtete Endpunkt hat zweifelsohne große Relevanz für die Bewertung des Nutzens einer diagnostischen oder therapeutischen Maßnahme. Aus der knappen Benennung dieses Endpunktes wird jedoch nicht ausreichend klar, wie das IQWiG diese Zielgröße in der Nutzenbewertung genau auffassen wird. In mehreren früheren Nutzenbewertungen hat das IQWiG diesen Punkt präziser gefasst. Als Beispiele sind nachfolgend einige solcher Bewertungsaufträge angeführt:

Bewertungsauftrag	**Festlegung der Zielgröße**
A05-04 (Kurzwirksame Insulinanaloga zur Behandlung des Diabetes mellitus Typ 2)	Erhalt bzw. Besserung krankheitsbezogener Lebensqualität
A05-13 (Fixe Kombinationen aus Kortikosteroiden und lang wirksamen Beta-2 Rezeptoragonisten zur inhalativen Anwendung bei Patienten mit Asthma bronchiale)	Besserung bzw. Erhalt der gesundheitsbezogenen Lebensqualität
A05-14 (Leukotrien-Rezeptor-Antagonisten bei Patienten mit Asthma bronchiale)	Besserung bzw. Erhalt der erkrankungsbezogenen Lebensqualität
A05-19A (Cholinesterasehemmer bei Alzheimer Demenz)	Besserung bzw. Erhalt der krankheitsbezogenen Lebensqualität; Besserung bzw. Erhalt der Lebensqualität der (betreuenden) Angehörigen

Bei einer chronisch progredienten Erkrankung wie der COPD nimmt die Lebensqualität im Verlauf von Monaten und Jahren im Mittel ab. Unter diesen Umständen ist auch von einer effektiven Therapie eine „Verbesserung“ der Lebensqualität (gegenüber dem Ausgangszustand bei Studienbeginn) nicht unbedingt zu erwarten. Je länger die Studiendauer ist, desto ausgeprägter wird der klinische Effekt einer Therapiemaßnahme von diesem nachteiligen Spontanverlauf der Erkrankung überlagert, sodass sich mit längerer Studiendauer das Therapieziel zunehmend von der „Verbesserung“ zum „Erhalt“ der Lebensqualität verlagert.

Daher bitten wir um eine Präzisierung der Zielgröße in der Weise, wie es auch in den o.g. anderen Bewertungsaufträgen erfolgt ist. Desweiteren sollte der Begriff „Lebensqualität“ durch den des „Gesundheitsstatus“ ergänzt werden, da die auf dem Gebiet der COPD gebräuchlichen Fragebogeninstrumente (wie z.B. der St. George’s Respiratory Questionnaire – SGRQ) genau genommen ein quantitatives Maß für den Gesundheitsstatus darstellen und nicht die Lebensqualität ermitteln[6].

Die Zielgröße „gesundheitsbezogene Lebensqualität“ sollte demnach folgendermaßen formuliert sein:
„Verbesserung bzw. Erhalt der gesundheitsbezogenen Lebensqualität / des Gesundheitsstatus“

2 Gegenüberstellung der Ergebnisse der Einzelstudien: Nichtberücksichtigung von Studienergebnissen, die auf weniger als 70 % der randomisierten Patienten beruhen

Unter Punkt 4.4.1 des vorläufigen Berichtsplans wird ausgeführt, dass Ergebnisse i. d. R. nicht in die Nutzenbewertung einfließen, „wenn diese auf weniger als 70% der in die Auswertung einzuschließenden Patienten basieren, d. h. wenn der Anteil der fehlenden Werte größer als 30% ist“. Diesbezüglich wird Bezug genommen auf die Arbeit von Schulz und Grimes[7], die einen Prozentsatz von 20 als oberste Grenze für einen noch tolerablen „loss to follow-up“ hinsichtlich der Validität einer Studie zitieren.

Insbesondere bitten wir um Klarstellung, dass sich die genannten Zahlen nicht auf die vorzeitigen Abbruchraten beziehen. Diese sind von der Länge der Studie abhängig und würden Langzeitstudien fast zwangsläufig invalide machen. Wir gehen vielmehr davon aus, dass sich die Zahlen auf den Anteil an Patienten beziehen, die nicht in die Auswertung eingeschlossen werden. Diesbezüglich bitten wir um Konkretisierung bzw. Präzisierung wie dies exakt zu verstehen ist, analog zur Anhörung des vorläufigen Berichtsplan zum Auftrag A09-01 (Dipyridamol + ASS zur Sekundärprävention nach Schlaganfall oder TIA). Das Problem von ggf. fehlenden Werten muss mit geeigneten statistischen Methoden adressiert werden.

Die Vorgabe im Vorläufigen Berichtsplan zum Bewertungsauftrag A05-18, dass Ergebnisse i. d. R. nicht in die Nutzenbewertung einfließen, „wenn diese auf weniger als 70% der in die Auswertung einzuschließenden Patienten basieren“, sollte dahingehend konkretisiert werden, dass sich dies auf den Anteil der Patienten, die in eine Analyse eingeschlossen werden, bezieht.

3 Informationssynthese und –analyse

Das IQWiG sieht im vorläufigen Berichtsplan die Bewertung der klinischen Relevanz eines beobachteten Patientennutzens vor. Dabei soll die Relevanzbewertung je nach Verfügbarkeit der Daten auf Basis von Mittelwertsdifferenzen und / oder anhand von Responderanalysen unter primärer Berücksichtigung von validierten MIDs (minimal important differences) erfolgen.

Hierbei wird jedoch nicht mitgeteilt, welche Relevanzgrenzen es pro Endpunkt gibt – und ob diese wohlbegründet sind. Dies gilt insbesondere für die Erfassung der Zielgrößen „COPD-Symptome“ und „gesundheitsbezogene Lebensqualität“. Bei COPD-Symptomen und bei der gesundheitsbezogenen Lebensqualität handelt es sich um subjektive Endpunkte, die in klinischen Studien in der Regel in Form von Fragebogen-basierten Skalen oder Scoring-Systemen erfasst werden. Zur Frage der Validität solcher Skalen oder Scoring-Systeme bzw. zum Vorliegen entsprechender Relevanzschwellen (minimal important differences – MIDs) hätte im Rahmen eines Scoping-Prozesses im Vorfeld der Berichtsplanerstellung eine Befragung von klinisch ausgewiesenen Experten der wissenschaftlich-medizinischen Fachgesellschaften im für die Fragestellung relevanten Therapiegebiet erfolgen sollen, wie dies in nationalen und internationalen Gutachten[1,2,6] wiederholt gefordert wurde.

Liegen für (komplexe) Skalen keine validierten Grenzen vor, so bezieht sich das IQWiG auf Hedges’ g von 0.2. Hedges’ g ist eine Modifikation von Cohen’s d. Es ist ein rein statistisches Maß, das zur Vereinfachung der Fallzahlabschätzung beschrieben wurde. Cohen bemerkt, dass mit der Einführung der Konvention, Effekte in klein, mittel und groß zu kategorisieren, nicht die Wichtigkeit (klinische Relevanz) von kleinen, mittleren oder großen Effekten präjudiziert ist. Er warnt vor einer Missdeutung und betont die Relativität der Kategorien in verschiedenen Anwendungsfeldern[3].

Die europäische Arzneimittelbehörde EMA lehnt bei der Bestimmung eines „non-inferiority margins“ solche Effektgrößen als alleinige Grundlage ab und fordert stattdessen eine Interpretation im klinischen Kontext: „It is not appropriate to use effect size (treatment difference divided by standard deviation) as justification for the choice of non-inferiority margin. This statistic provides information on how difficult a difference would be to detect, but does not help justify the clinical relevance of the difference, and does not ensure that the test product is superior to placebo.“[4]

Das IQWiG will – falls Responderauswertungen nicht zur Verfügung stehen – einen Test mit verschobenen Nullhypothesen anwenden, indem es fordert, dass das zum beobachteten Effekt korrespondierende Konfidenzintervall vollständig oberhalb der Relevanzschwelle liegt, damit von einer relevanten Effektstärke ausgegangen werden kann. Dieser Test ist bei klinischen Prüfungen, außer im Nicht-Unterlegenheits-Design, unüblich. Bei dieser Vorgehensweise ist der Nachweis, dass eine MID statistisch signifikant überschritten wird, nur mit sehr großen Fallzahlen zu führen. Die MID, die durchaus relevant ist, kann praktisch nicht nachgewiesen werden.

Wir halten die Einbeziehung von Fachgesellschaften im Rahmen eines Scoping-Prozesses zur Frage der Beurteilung der klinischen Relevanz eines Patientennutzens für sehr wichtig. Der Wert der Effektgröße (Hedges' g) von 0,2, den das IQWiG als Minimalwert einer klinischen Relevanz für alle Messskalen und Patientenpopulationen ansieht, stellt aus unserer Sicht keinen international verbindlichen Richtwert dar. Vielmehr ist dieser Ansatz rein technischer Natur und soll mit einem statistischen Test eine binäre Entscheidung treffen. Letztlich ist aber jede statistisch signifikante Verbesserung eines patientenrelevanten Endpunkts als Zusatznutzen aus klinischer Perspektive positiv zu bewerten.

4 Literaturverzeichnis:

1. Antes G, Jöckel KH, Kohlmann T, Raspe H, Wasem J. Kommentierende Synopse der Fachpositionen zur Kosten-Nutzenbewertung von Arzneimitteln – Erstellt im Auftrag des Bundesministeriums für Gesundheit 2007
2. Bekkering GE, Kleijnen J. Procedures and methods of benefit assessments for medicines in Germany. Eur J Health Econ 2008; 9 (Suppl 1): 5-29 bzw. Dtsch Med Wochenschr 2008; 133 (Suppl 7): S225-S246
3. Cohen J. Statistical power analysis for the behavioral sciences, 2nd Edition. Taylor & Francis Group, New York 1988
4. EMA – Committee for medicinal products for human use (CHMP). Guideline on the choice of inferiority margin 2005
5. Jones PW. Health status and the spiral of decline. COPD 2009; 6: 59-63
6. NICE – National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal. http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf (letzter Zugriff: 17.05.2010)
7. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet 2002; 359: 781-785

A 1.2 Deutsche Atemwegsliga e. V.

Autor:
Kardos, Peter, Dr.

Adresse:
Deutsche Atemwegsliga e. V.
Im Prinzenpalais: Burgstraße
33175 Bad Lippspringe

DEUTSCHE ATEMWEGSLIGA E.V.

Stellungnahme zum Vorläufigen Berichtsplan A05-18 Tiotropiumbromid bei COPD

Die Vorstände der Deutschen Atemwegsliga e.V. (DAL) und der Deutschen Gesellschaft für Pneumologie und Beatmungsmedizin (DGP) haben beschlossen, zu den nachfolgend aufgeführten Punkten im o.g. Vorbericht Stellung zu nehmen. Tiotropium ist derzeit ein mit dem höchst möglichen Evidenzgrad belegtes Standardmedikament in der Dauertherapie von COPD Patienten. Es spielt auch eine zentrale Rolle in der internationalen (GOLD-, http://www.goldcopd.com/) und in der nationalen, von DAL und DGP herausgegebenen Leitlinie [1] ebenso, wie in der nationalen Versorgungsleitlinie COPD http://www.versorgungsleitlinien.de/themen/copd/pdf/nvl_copd_lang.pdf..

Die geplante Nutzenbewertung von Tiotropiumbromid im Vergleich zu einer Placebogabe oder anderen medikamentösen Therapieoptionen, einzeln oder in Kombination, sollte den klinischen Wert des Medikamentes in der Dauertherapie abbilden.

Der geplante Ausschluss von einzelnen Endpunkten von Studien, mit einer drop out Rate von mehr als 30% bezogen auf den jeweiligen Endpunkt beinhaltet ein erhebliches Verzerrungspotential für die oben genannte Zielsetzung. Wichtige Endpunkte vieler für die COPD Therapie bedeutenden, in sehr angesehenen Zeitschriften (British Medical Journal, New England Journal of Medicine, American Journal of Respiratory and Critical Care Medicine) publizierten Langzeitstudien [2-5] würden durch dieses Kriterium aus der Nutzenbewertung ausgeschlossen werden. Bei zweier dieser als Meilenstein erachteten Studien [4;5] wurde Tiotropium evaluiert.
Je länger die Dauer einer Studie ist, umso niedriger ist die Anzahl der am Ende der Studien verbliebenen Patienten. (Tab. 1). Gerade die Langzeitstudien bilden aber am besten den Nutzen eines für die jahrelange Dauertherapie der COPD indizierten Medikamentes ab.
In der UPLIFT Studie [4], dessen wichtige Endpunkte nach dem Berichtsplan aus der Bewertung ausgeschlossen werden sollten, wurde die erwartete Studienabbruchrate für die co-primären Endpunkte mit überraschend zutreffender Genauigkeit im Voraus statistisch berücksichtigt. Die Patientenzahl sei laut Publikation so kalkuliert worden, dass bei 40% der eingeschlossenen Patienten noch valide prospektive Subgruppenanalysen durchgeführt werden können.
Die statistische Auswertung verschiedener Endpunkte bei COPD Patienten ist Thema kontroverser Diskussionen, derzeit gibt es noch keinen allgemein akzeptierten Goldstandard [6-11].
Die zur Begründung des Ausschlusses von IQWIG zitierte Arbeit von Kenneth F. Schulz aus dem Lancet 2002 zitiert lediglich eine Empfehlung von Sackett aus dem Jahre 2000. Danach sei ein „lost to follow up" von mehr als 20% inakzeptabel. Schulz bezieht sich nicht auf die gesamte Abbruchrate. Studienabbrüche resultieren nicht nur aus „lost to follow up". Andere

Gründe, zum Beispiel der Widerruf der Einwilligung (immerhin 10% in den 4 Jahren der Uplift Studie [4]) tragen ebenfalls dazu bei. Der Ausschluss von Patienten mit Widerruf der Einwilligung erlaubt dennoch eine ITT Analyse. Die 20% Rate ist darüber hinaus eine allgemeine Empfehlung und sie wird den speziellen Gegebenheiten einer Langzeitstudie nicht gerecht. Das erst kürzlich publizierte Update der CONSORT Leitlinie [12] berücksichtigt bereits die praktischen Schwierigkeiten bei der statistischen Behandlung der Studienabbrecher: „Like "intention-to-treat," none of these other labels reliably clarifies exactly which patients were included. Thus, in the CONSORT checklist we have dropped the specific request for intention-to-treat analysis in favour of a clear description of exactly who was included in each analysis".
Wir sind der Meinung, dass die Festlegung einer Abbruchrate die zum Ausschluss von Endpunkten großer, in angesehenen Zeitschriften publizierter Studien führt, nicht gerechtfertigt ist. Dieses Verfahren könnte zu Aussagen führen, die das Krankheitsbild COPD nicht adäquat abbilden.
Bei COPD handelt es sich um eine langsam progrediente Erkrankung (s. Definition auf Seite 1 des Berichtplans). Von besonderem Wert für die Nutzenbewertung sollten daher Studien mit zwei, drei und mehr Jahren Studiendauer sein. Lungenfunktionsergebnisse oder körperliche Belastbarkeit lassen sich in kürzeren Studien valide bewerten. Gerade für die im Berichtsplan vorgesehenen patientenrelevanten Endpunkte (Exazerbationsfrequenz, Mortalität, Krankenhausbehandlungen, COPD-assoziierte kardiovaskuläre Morbidität und Mortalität, COPD-bedingte Letalität und Gesamtmortalität, unerwünschte Arzneimittelwirkungen) sind Halbjahresstudien wegen prominenter saisonaler Unterschiede und aus statistischen Überlegungen (zu kleine Anzahl der eingetretenen Ereignisse) mit erheblichem Verzerrungspotential (bias) behaftet. Selbst Einjahresstudien dürften aufgrund der Seltenheit der Ereignisse für die Bewertung der Mortalität, der unerwünschten Arzneimittelwirkungen weniger geeignet sein als die Langzeitstudien, die nun nicht oder nur partiell bewertet werden sollen.

In die Nutzenbewertung sollten nach dem Berichtsplan nur Studien einfließen, die einen Vergleich zu Placebo oder zu anderen medikamentösen Therapieoptionen prüfen. Eine kausale Behandlung COPD (abgesehen von der Raucherentwöhnung) ist nicht vorhanden; nichtmedikamentöse Maßnahmen wie die Rehabilitation spielen daher eine besonders wichtige Rolle. Wir möchten deshalb anregen, dass auch Studien gewertet werden, in denen der Zusatznutzen von Tiotropium in Rahmen einer Rehabilitation dargestellt wird [13]. Nur wenige Untersuchungen liegen vor, die den Zusatznutzen eines Pharmakons bei der Rehabilitation von COPD Patienten auswerten.

Desweiteren empfehlen wir den Einschluss von FEV_1 als Endpunkt, da die FEV_1 der am besten validierte Surrogatparameter für Mortalität und Exazerbationen darstellt. [14;15].

Die Atemwegliga und die DGP sind überzeugt, dass eine objektive Bewertung des therapeutischen Nutzens von Tiotropium nur dann erfolgen kann, wenn Änderungen am vorgelegten vorläufigen Berichtsplan vorgenommen werden.

Deutsche Atemwegsliga e.V., Geschäftsstelle. Im Prinzenpalais/Burgstraße, 33175 Bad Lippspringe

Tel.: 05252/933615 Fax: 05252/933616, E-Mail: atemwegsliga.lippspringe@t-online.de

Tabelle 1 Abbruchraten in einigen Langzeitstudien bei COPD Patienten

Autor [Literaturstelle]	Journal	Studiendauer	Abbruchrate (%) Tiotropium	Abbruchrate (%) Prüfsubstanz
Brusasco [16]	Thorax 2003	6 Monate	15,4	
Casaburi [13]	Chest 2005	6 Monate	14.5	
Donohue [17]	Chest 2002	6 Monate	12	
Niewoehner [18]	Ann Intern Med 2005	6 Monate	16	
Vogelmeier [19]	Respir Med 2008	6 Monate	13	
Tonnel [20]	In J of COPD 2008	9 Monate	14.7	
Chan [21]	Can Respir J 2007	48 Wochen	22.2	
Casaburi [22]	Eur Respir J 2002	1 Jahr	18.7	
Dusser [23]	Eur Respir J 2006	1 Jahr	13.4	
Powrie [24]	Eur Respir J 2007	1 Jahr	30.4	
Vincken [25]	Eur Respir J 2002	1 Jahr	15.2	
Wedzicha [5]	Am J Respir Crit Care Med 2008	2 Jahre	35	
Burge [2]	Thorax 2003	3 Jahre		43%
Calverley [3]	N Engl J Med 2007	3 Jahre		34%
Tashkin [4]	N Engl J Med 2008	4 Jahre	36%	

Frankfurt am Main, den 15.5.2010

In Vertretung:

Dr. med. Peter Kardos

Gemeinschaftspraxis und Zentrum Allergologie, Pneumologie, Schlafmedizin

Klinik Maingau, Scheffelstrasse 33

60318 Frankfurt am Main

Literatur

1. Vogelmeier C, Buhl R, Criee CP et al. Leitlinie der Deutschen Atemwegsliga and der Deutschen Gesellschaft für Pneumologie und Beatmungsmedizin zur Diagnostik und Therapie von Patienten mit chronisch obstruktiver Bronchitis und Lungenemphysem (COPD). *Pneumologie* 2007; 61:e1-40

2. Burge PS, Calverley PM, Jones PW et al. Prednisolone response in patients with chronic obstructive pulmonary disease: results from the ISOLDE study. *Thorax* 2003; 58:654-658

3. Calverley PM, Anderson JA, Celli B et al. Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *N Engl J Med* 2007; 356:775-789

4. Tashkin DP, Celli B, Senn S et al. A 4-year trial of tiotropium in chronic obstructive pulmonary disease. *N Engl J Med* 2008; 359:1543-1554

5. Wedzicha JA, Calverley PM, Seemungal TA et al. The prevention of chronic obstructive pulmonary disease exacerbations by salmeterol/fluticasone propionate or tiotropium bromide. *Am J Respir Crit Care Med* 2008; 177:19-26

6. Aaron SD, Fergusson D, Marks GB et al. Counting, analysing and reporting exacerbations of COPD in randomised controlled trials. *Thorax* 2008; 63:122-128

7. Keene ON, Vestbo J, Anderson JA et al. Methods for therapeutic trials in COPD: lessons from the TORCH trial. *Eur Respir J* 2009; 34:1018-1023

8. Kardos P Methodological issues in therapeutic trials of COPD. *Eur Respir J* 2009; 33:443-444

9. Keene ON Intent-to-treat analysis in the presence of off-treatment or missing data. *Pharm.Stat.* 2010;

10. Suissa S Statistical Treatment of Exacerbations in Therapeutic Trials of Chronic Obstructive Pulmonary Disease. *Am.J.Respir.Crit.Care Med.* 2006; 173:842-846

11. Suissa S Methodologic Shortcomings of the INSPIRE Randomized Trial. *Am.J.Respir.Crit.Care Med.* 2008; 178:1090-109b

Deutsche Atemwegsliga e.V., Geschäftsstelle. Im Prinzenpalais/Burgstraße, 33175 Bad Lippspringe

Tel.: 05252/933615 Fax: 05252/933616, E-Mail: atemwegsliga.lippspringe@t-online.de

12. Moher D, Hopewell S, Schulz KF et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 340:c869

13. Casaburi R, Kukafka D, Cooper CB et al. Improvement in Exercise Tolerance With the Combination of Tiotropium and Pulmonary Rehabilitation in Patients With COPD. *Chest* 2005; 127:809-817

14. Celli BR, Cote CG, Marin JM et al. The Body-Mass Index, Airflow Obstruction, Dyspnea, and Exercise Capacity Index in Chronic Obstructive Pulmonary Disease. *N Engl J Med* 2004; 350:1005-1012

15. Donaldson GC, Wedzicha JA COPD exacerbations .1: Epidemiology. *Thorax* 2006; 61:164-168

16. Brusasco V, Hodder R, Miravitlles M et al. Health outcomes following treatment for six months with once daily tiotropium compared with twice daily salmeterol in patients with COPD. *Thorax* 2003; 58:399-404

17. Donohue JF, van Noord JA, Bateman ED et al. A 6-month, placebo-controlled study comparing lung function and health status changes in COPD patients treated with tiotropium or salmeterol. *Chest* 2002; 122:47-55

18. Niewoehner DE, Rice K, Cote C et al. Prevention of exacerbations of chronic obstructive pulmonary disease with tiotropium, a once-daily inhaled anticholinergic bronchodilator: a randomized trial. *Ann Intern Med* 2005; 143:317-326

19. Vogelmeier C, Kardos P, Harari S et al. Formoterol mono- and combination therapy with tiotropium in patients with COPD: a 6-month study. *Respir Med* 2008; 102:1511-1520

20. Tonnel AB, Perez T, Grosbois JM et al. Effect of tiotropium on health-related quality of life as a primary efficacy endpoint in COPD. *Int.J Chron.Obstruct.Pulmon.Dis.* 2008; 3:301-310

21. Chan CK, Maltais F, Sigouin C et al. A randomized controlled trial to assess the efficacy of tiotropium in Canadian patients with chronic obstructive pulmonary disease. *Can.Respir J* 2007; 14:465-472

22. Casaburi R, Mahler DA, Jones PW et al. A long-term evaluation of once-daily inhaled tiotropium in chronic obstructive pulmonary disease. *Eur.Respir.J.* 2002; 19:217-224

Deutsche Atemwegsliga e.V., Geschäftsstelle. Im Prinzenpalais/Burgstraße, 33175 Bad Lippspringe

Tel.: 05252/933615 Fax: 05252/933616, E-Mail: atemwegsliga.lippspringe@t-online.de

23. Dusser D, Bravo ML, Iacono P et al. The effect of tiotropium on exacerbations and airflow in patients with COPD. *Eur Respir J* 2006; 27:547-555

24. Powrie DJ, Wilkinson TMA, Donaldson GC et al. Effect of tiotropium on sputum and serum inflammatory markers and exacerbations in COPD. *Eur Respir J* 2007; 30:472-478

25. Vincken W, van Noord JA, Greefhorst AP et al. Improved health outcomes in patients with COPD during 1 yr's treatment with tiotropium. *Eur.Respir.J.* 2002; 19:209-216

Deutsche Atemwegsliga e.V., Geschäftsstelle. Im Prinzenpalais/Burgstraße, 33175 Bad Lippspringe

Tel.: 05252/933615 Fax: 05252/933616, E-Mail: atemwegsliga.lippspringe@t-online.de

A 1.3 GlaxoSmithKline GmbH & Co. KG

Autoren:
Lützelberger, Uwe
Banik, Norbert, Dr. Dr.
Hagedorn, Cordula, Dr.

Adresse:
GlaxoSmithKline GmbH & Co. KG
Theresienhöhe 11
80339 München

Stellungnahme der GlaxoSmithKline GmbH & Co. KG zum vorläufigen Berichtsplan „Tiotropium bei COPD“

(Auftrag A05-18, Version 1.0 vom 12.04.2010)

Anmerkungen zu der im vorläufigen Berichtsplan (Version 1.0) dargestellten projektspezifischen Methodik:

1. Stellenwert der Inhalationssysteme in der Bewertung

Im Kapitel 1 (S. 5) des vorläufigen Berichtsplans wird auf die Inhalationssysteme eingegangen. Es werden beispielhaft Kriterien genannt, die die Wahl/Empfehlung des oder der Inhalationssysteme beeinflussen können. Diese Aufzählung ließe sich ergänzen um z.B. Faktoren der unmittelbaren Deposition der inhalierten Substanz(en) in der Lunge und des Einflusses des Inhalationssystems auf die Adhärenz oder Compliance des Patienten.
Die mögliche Bedeutung des Inhalationssystems für die Nutzenbewertung unabhängig oder zusätzlich zum applizierten Wirkstoff selbst wird im zweiten Teilziel dieser Bewertung für Tiotropiumbromid hervorgehoben und im zweiten Abschnitt zum Inhalationssystem (S.6) detailliert ausgeführt.

Unseres Erachtens ist jedoch im vorläufigen Berichtsplan das Vorgehen bei unterschiedlichen Inhalationssystemen in den Vergleichstherapieoptionen bisher nicht berücksichtigt worden. Wir bitten deshalb darum, an geeigneter Stelle im Berichtsplan darauf einzugehen und darzulegen, wie im Detail unterschiedliche Inhalationssysteme der Vergleichstherapien bei der Nutzenbewertung berücksichtigt werden. Wir sind der Meinung, dass der Faktor „Inhalationssystem“ – falls für einzelne Bewertungen relevant - zu berücksichtigen ist und zumindest im Sinne von Subgruppenanalysen (Kapitel 4.4.4) einfließen sollte. Inwieweit sich weitere Konsequenzen aus der durchgehenden Berücksichtigung des Faktors „Inhalationssysteme“ für die Informationssynthese- und Analyse (4.4) oder andere Punkte des Berichtsplans ergeben, sollte bei der Überarbeitung des vorläufigen Berichtsplans ebenfalls geprüft und ggfs. einbezogen werden.

2. Genauere Spezifikation von „Asthmapatienten“ und „Patienten mit COPD“

Es erscheint uns erforderlich, für diesen Berichtsplan (speziell in 4.1.1) genauer zu spezifizieren, wie der Begriff „Asthmapatienten“ in der Nutzenbewertung verstanden und operationalisiert werden soll.

Im Umkehrschluss ist es ebenfalls erforderlich „Patienten mit COPD“ genauer zu definieren bezüglich der Kriterien, die für diese Diagnose angelegt werden sollen. In Kapitel 1 werden dazu zwar generelle, unseres Erachtens aber keine hinreichend expliziten Angaben gegeben.

Eine genaue Definition ist aber bereits an dieser Stelle des Berichtsplans erforderlich, um die Kriterien der Studienselektion genau festlegen und darauf aufbauend auch die Kriterien aus Abschnitt 4.1.8 zuverlässig beurteilen zu können.

3. Klärung des Umfangs der Nutzenbewertung bezüglich der Tiotropium-Behandlung und der Vergleichsbehandlungen hinsichtlich zu berücksichtigender Kombinationen mit weiteren Wirkstoffen

Als erstes Ziel der Untersuchung werden unter Punkt 2 die „Nutzenbewertung von Tiotropiumbromid im Vergleich zu einer Placebogabe oder anderen medikamentösen Therapieoption, einzeln oder in Kombination …“ genannt. Welche Medikamentenkombinationen hier in welchem Umfang berücksichtigt werden, ist an dieser Stelle jedoch nicht ausgeführt. Diese Anmerkung gilt auch für den Unterpunkt 4.1.2 „Prüf- und Vergleichsintervention“ unter a): „Die zu prüfende Intervention ist die inhalative Dauertherapie der COPD mit Tiotropiumbromid. Als Vergleichsintervention wird eine Placebobehandlung oder eine andere in Deutschland verfügbare medikamentöse Therapieoption (einzeln oder in Kombination) zur inhalativen dauertherapie der COPD betrachtet.“. Gleiches gilt sinngemäß für den Abschnitt 4.1.6, Tabelle 3 (E2a/E3a), in dem dort auf 4.1.2 verwiesen wird.

Für die medikamentöse Behandlung der COPD stehen sowohl Monotherapien als auch Medikamentenkombinationen zur Verfügung. Der Stufenplan für Prophylaxe und Langzeittherapie, der auf der Leitlinie basiert, ist in Abschnitt 1, Tabelle 2, veranschaulicht worden. Demnach kann Tiotropiumbromid als inhalative Therapie mit einem langwirksamen antimuscarinergen Wirkstoff (LAMA) prinzipiell ab dem Schweregrad II eingesetzt werden.

Aus dem Stufenplan ergibt sich damit, dass Tiotropiumbromid als Monotherapie oder in Kombination mit z.B. einem kurzwirksamen Bronchodilatator (SABA) nach Bedarf und/oder in Kombination mit anderen Wirkstoffen eingesetzt werden kann. Relevant sind hierbei andere langwirksame Bronchodilatatoren (LABA) und inhalative Corticosteroide (ICS) sowie wiederum Kombinationen dieser Wirkstoffe.

Unseres Erachtens geht aus dem vorläufigen Berichtsplan nicht eindeutig hervor, ob und wenn ja welche Kombinationen von Tiotropiumbromid mit anderen relevanten Wirkstoffen und folglich welche Vergleichsbehandlungen konkret in die Nutzenbewertung einbezogen werden. Wir bitten deshalb darum, im Berichtsplan, entsprechend unserer obigen Ausführungen im Detail darzulegen, welche Therapieoptionen in die Nutzenbewertung eingehen sollen. Die Relevanz dieser Fragestellung spiegelt sich auch darin wieder, dass zu zahlreichen dieser medikamentösen Kombinationstherapien (mit bzw. ohne Tiotropiumbromid) sowohl Ergebnisse von randomsierten Einzelstudien als auch von systematischen Reviews vorliegen.

4. Bearbeitung von Teilziel a) (Tiotropiumbromid versus Komparatoren) in Abhängigkeit vom Ausgang von Teilziel b) (Vergleich von HandiHaler versus Respimat)

Zur weiteren Spezifikation der Bearbeitung von Teilziel a) erscheint es uns wichtig (in Erweiterung der Ausführungen zu 4.1.2) im Berichtsplan anzugeben, wie die Bearbeitung von Teilziel a) im Detail erfolgen soll, falls sich im Ergebnis der Bearbeitung des Teilziels b) eine Differenzierung der beiden verfügbaren Inhalatoren des Tiotropiumbromids bezüglich Nutzens- oder Schadensaspekten oder des Nutzen-Schaden-Verhältnisses ergibt.
In einem solchen Fall sollte die zu prüfende Intervention für Teilziel a) ebenfalls getrennt (stratifiziert) nach den beiden verfügbaren Inhalatoren für Tiotropiumbromid mit den Vergleichstherapien verglichen und bewertet werden (soweit Evidenz dazu vorliegt).

5. Verwendung von nicht randomisierten Studien zusätzlich zu RCT

Unter 4.1.4 wird erneut darauf verwiesen, dass lediglich RCT in die Nutzenbewertung einfließen sollen. Wir sind nach wie vor der Meinung, dass das Prinzip der bestverfügbaren Evidenz hier anzuwenden ist und einer Beurteilung des Nutzens einer Intervention, die sich unter Routineanwendung befindet, wesentlich besser gerecht wird. Entscheidungen in diesem Kontext sind immer mit Unsicherheit behaftet und es ist deshalb inhärenter und

konsequenter Bestandteil einer jeden (Nutzen-) Bewertungsaufgabe, den Grad der Unsicherheit zu beurteilen und in die Schlussfolgerungen einer Bewertung einfließen zu lassen. Am Beispiel der RCT wird dies ja auch bereits seit längerem vom IQWiG so umgesetzt, in dem separat das Verzerrungspotential der RCT beurteilt wird. Mit GRADE liegt zudem ein erprobtes und bewährtes Instrument für eine differenzierte Bewertung von Gesamtevidenz unterschiedlicher Evidenzstufen vor [1], das auch von vergleichbaren HTA-Instituten (z.B. dem NICE) oder von der Cochrane Collaboration, der WHO und der American Thoracic Society (ATS) eingesetzt wird.

Wir regen deshalb neuerlich und nachdrücklich an, bei der Studienselektion nicht von vornherein ausschließlich auf RCT zu rekurrieren. Auch der Verweis darauf, dass diese möglich und praktisch durchführbar seien hilft dabei nicht. Das Ausschlaggebende ist doch, dass diese Studien vorliegen müssen, um einbezogen werden zu können, nicht, dass sie vorliegen könnten.

Es erscheint dem Kosten-Nutzen-Verhältnis des Erkenntnisgewinns nicht angemessen, im Falle des Nicht-Vorliegens von RCT's, vorhandene relevante nicht-randomisierte Studien komplett und a priori geplant zu ignorieren. Speziell – aber nicht darauf beschränkt – bei der Beurteilung von Schadensmerkmalen von Interventionen kann durch Einbeziehen von nicht-randomisierten Studien zusätzliche Evidenz zu einer Gesamtbeurteilung beigetragen werden – entweder in Ergänzung zu RCT oder - bei deren Nichtverfügbarkeit - als alleinige Evidenzbasis.

6. Studiendauer

Wir bitten zu überdenken, ob es wirklich im Sinne der Abklärung der patientenrelevanten Endpunkte ausreichend ist, ausschließlich Studien mit einer Mindestdauer von 6 Monaten (24 Wochen) zu berücksichtigen. Zwar ist die Anlehnung an die Guideline der EMA (Referenz Nr. 17 des vorläufigen Berichtsplans) prinzipiell natürlich naheliegend, jedoch wird auch dort für das Zulassungsverfahren lediglich gefordert, dass der Zulassungsantrag für die Indikation COPD mit einer Studie mit mindestens 6 Monaten Dauer unterstützt werden sollte, nicht also, dass alle Studien diese Dauer aufweisen müssen. Hierbei und auch in der Behandlungssituation bedeutet dies, dass auch in kurzfristigeren Studien patientenrelevante Effekte auftreten und nachgewiesen werden können (z.B. Lebensqualität; Exazerbationshäufigkeiten und/oder Häufigkeit von Krankenhauseinweisungen in entsprechend schwergradigen Patientenpopulationen). Solange nicht Evidenz dafür vorliegt, dass solche Effekte transient sind, ist ein patientenrelevanter Effekt auch über 12 Wochen in einer Nutzenbewertung nicht zu negieren. Es könnte vorkommen, dass für bestimmte Konstellationen aus Vergleichsmedikation(en) und Inhalationssystemen einzelne

patientenrelevante Endpunkte nur in Studien mit einer Maximaldauer von beispielsweise 12 Wochen Behandlung untersucht wurden.
Wir sind der Meinung, dass die EMA in ihrer Guideline keineswegs ausschließlich Studien mit einer Mindestdauer von 6 Monaten gefordert hat und dass unter Hinblick auch auf eher schwere Stadien der COPD relevante Effekte auch bei einem Nachweis von über 12 Wochen Behandlungsdauer berücksichtigt werden sollten, da sie für Patienten wahrnehmbar und damit relevant sein können. Deshalb sollte die Mindeststudiendauer bei der Literaturrecherche auf 12 Wochen herabgesetzt werden und dann in der Nutzenbewertung identifiziert werden, welche Studien in welcher Vergleichssituation zusätzlich zu den Studien mit mindestens 6 Monaten Behandlungsdauer Erkenntnisse zu patientenrelevanten Endpunkten liefern können – und zu welchen Endpunkten.

7. Relevanzbetrachtung

In 2009 hat das IQWiG in Vorberichten zu Nutzenbewertungen Bewertungen zur Relevanz von Studienergebnissen kontinuierlicher patientenrelevanter Variabler basierend auf einem Cohen's d von mindestens 0.2 eingeführt (A05-20A und A05-19C). Dieses damals ungeplante Vorgehen hat in jenen konkreten Bewertungsverfahren und darüber hinaus zu einer lebhaften wissenschaftlichen Debatte geführt, die nicht abgeschlossen ist. So wird sich auch das IQWiG bei seiner diesjährigen Veranstaltung „IQWiG im Dialog 2010" dieses Themas annehmen.
GSK hat bei seiner schriftlichen Stellungnahme und in der mündlichen Erörterung zum Vorbericht A05-20C ebenfalls Argumente mit dem IQWiG diskutiert. Wir mussten jedoch im Abschlussbericht zu A05-20C konstatieren, dass das IQWiG bezüglich des Kriteriums selbst und dessen Operationalisierung keinerlei Änderung vorgenommen hat.

Angeregt u.a. durch die o.g. wissenschaftliche Debatte möchten wir hier einige neue Punkte vorbringen, die unseres Erachtens für diesen konkreten Auftrag und für die generelle methodische Herangehensweise einer Relevanzbeurteilung bei Nutzenbewertungen im Rahmen der Bewertungsverfahren des IQWiG von Bedeutung sind. Wir hoffen, dass in einer wissenschaftlichen Debatte diesbezüglich Fragen geklärt werden können und in der Folge Relevanzbewertungen in diesem Verfahren und in Zukunft wissenschaftlich konsistent, im Gesamtkontext aus Nutzenbewertung und Kosten-Nutzenbewertung, erfolgen.

a) Generelle Eignung eines standardisierten Effektschätzers als Kriterium der (medizinischen, patientenbezogenen) Relevanz:

Die Debatte des vergangenen Jahres hat dazu geführt, dass das IQWiG selbst mittlerweile anerkennt, dass verteilungsbasierte Verfahren allein kein direktes und belastbares Korrelat zur „Relevanz" besitzen. Im Abschlussbericht A05-20C, Version 1.0 vom 09.11.2009, heißt es auf S. 434 dazu: „Es ist richtig, dass es sich bei Cohen's d, einem statistischen Maß, lediglich um eine Näherung an das Konzept der Relevanz handeln kann." und weiter „Cohen's d, als Maß der Effektstärke zur Beschreibung von Gruppenunterschieden, ist ein international häufig verwendeter Parameter zur Interpretation von Studienergebnissen…".

Es scheint uns wichtig, dies hervorzuheben: Cohen's d (und alle verteilungsbasierten Verfahren wie es standardisierte Mittelwertdifferenzen (SMD's) sind, also auch z. B. Hedge's g) sind an sich, d.h. ohne zusätzliche inhaltliche „Anker" kein Relevanzkriterium.
Zwei beispielhafte Zitate unterstreichen diese Einschätzung:

- „The distribution-based indices provide no *direct* information about the MID (minimum important difference). They are simply a way of expressing the observed change in a standardized metric. This makes it possible to compare change observed for measures that have a different raw metric and the degree of deviation (individual and group level) within the sample. ES (effect size) estimates can be compared to Cohen's guidelienes about the magnitude, but anchor based methods are the only way to estimate the MID directly." [2]; Erklärungen der Abkürzungen in runden Klammern von GSK
- "The relative improvement ((P – D)/P) and standardized difference or effect size ((P – D)/S) were used only to put things on a relative basis, not to quantify importance. Cohen's rule of small, moderate or large effect sizes (0.2, 0.5 and 0.8) was never found to be useful." [3]

Auch die „international häufige Verwendung", die wir so nicht verifizieren können, wäre per se kein valides Argument, da auch fehlerhafte Interpretationen von p-Werten und Konfidenzintervallen international sehr weit verbreitet sind, dadurch aber nicht richtig werden.

Wichtiger jedoch ist, dass die dazu zitierte Literaturstelle ([207] im o.g. Abschlußbericht – das „Cochrane Handbook for Systematic Reviews of Interventions", Kapitel 17 – SMD's lediglich in Bezug auf deren Wert erwähnt, unterschiedliche Endpunkte vergleichbar beurteilen zu können, nicht aber um die Relevanz von Endpunkten zu beurteilen. Folgerichtig findet man den Ansatz, SMD's für Relevanzaussagen zu benutzen, bei Systematic Reviews (SR's) der Cochrane Collaboration und in anderen publizierten SR's auch sehr selten. In Kapitel 12 des zitierten „Cochrane Handbook for Systematic Reviews of Interventions" [4] werden die

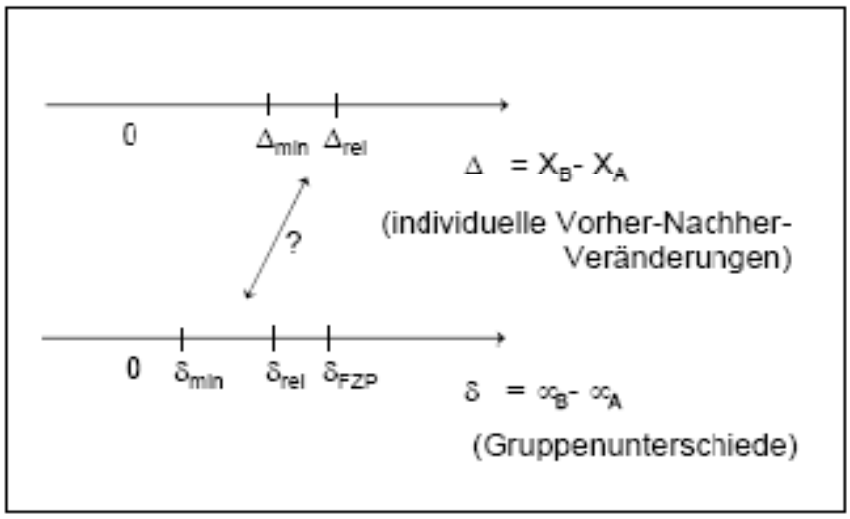
überwiegenden Schwierigkeiten einer Interpretation von SMD's im Sinne von „Relevanz" durch Ausführungen darüber bekräftigt, wie durch Umwandlungen von SMD's in Odds Ratios (OR) oder NNT's dem Manko der fehlenden Interpretierbarkeit der SMD's als Relevanzkriterium abgeholfen werden könnte. An dieser Stelle kann man erkennen, wie die Kontextabhängigkeit solcher Maße einer verlässlichen Relevanz-Interpretation von SMD's zuwiderläuft und keinesfalls nur SMD's von ≥ 0.2 einen „relevanten" Effekt verkörpern. Eine „Übersetzung" von SMD's in NNT's führt beispielsweise bei einer SMD = 0.1 (d.h. Cohen's d = 0.1) und einer „Responserate" in der Kontrollgruppe von 30% zu einer fast gleichen NNT von 27 wie bei einer SMD = 0.2 (Cohen's d = 0.2) und einer Responserate in der Kontrollgruppe von 10% [Ref. 207, Kapitel 12]. Es sind also nicht nur SMD's von mindestens 0.2 als relevant anzuerkennen; auch kleinere Werte können für einen ähnlichen „Impact" stehen: Es kommt auf den Kontext an.

b) Interpretation der Relevanz über die Konfidenzintervall-Ausschlussregel bzw. Testen mit verschobener Nullhypothese; Wahl des Kriteriums für dieses Vorgehen

Das IQWiG bewertet einen Effekt, der statistisch signifikant ist, nur dann als relevant, wenn zusätzlich das standardisierte Effektmaß (SMD), der sogenannte „Relevanzschätzer" (bisher Cohen's d jetzt in dieser Bewertung Hedge's g) statistisch signifikant größer als die vom IQWiG gewählte Relevanzschwelle g = 0.2 ist. Dieses Konzept entspricht einem Hypothesentest mit einer um die Relevanzschwelle (g oder d = 0.2) verschobenen Nullhypothese (statistical test with shifted null hypothesis).

In unserer bereits zitierten Stellungnahme zu A05-20C und in der darauf folgenden mündlichen Erörterung hatten wir auf die Implikationen eines solchen Vorgehens bereits hingewiesen und in Frage gestellt, ob eine damit de facto 97,5-prozentige Konfidenz, dass eine gefundene SMD größer als 0.2 ist, tatsächlich eine angemessene und post-hoc erfüllbare Forderung ist. Diesen Überlegungen hatte sich das IQWiG nicht anschließen können sondern geschrieben: „Der Vorschlag [ein niedrigeres Konfidenzniveau für dieses Entscheidungsproblem anzuwenden, Anmerkung GSK] bedeutet deshalb, dass die Relevanzgrenze bei Beibehaltung des üblichen Irrtumsniveaus abgesenkt würde. Eine Absenkung der Relevanzgrenze unter den Wert von 0.2 hält das IQWiG nicht für gerechtfertigt." [Abschlußbericht A05-20C, S. 437].
Diese Argumentation zeigt uns, dass der vorher selbst zugestandene sehr eingeschränkte Wert einer SMD als „Relevanzmaß" leider nicht konsequent umgesetzt wird und dass, zweitens, das Aufstellen einer „Relevanzgrenze" basierend auf einem festen Wert von z. B. 0.2 nicht fundiert ist.

Was bei dieser Interpretation jedoch nicht gesehen wird, ist die sinnvolle und inhaltlich konsistente Wahl von sinnvollen Schwellenwerten für ein solches Entscheidungsproblem. Wie Thomas [5] unter Bezug auf Victor [6] darstellt, ist es plausibel, unterschiedliche Schwellenwerte für Unterschiede auf individuellem Niveau und für Gruppenunterschiede zu betrachten und dann nochmals zu differenzieren zwischen einem „wünschenswerten und für eine Patientenpopulation nicht zu übersehenden Effekt“ (δ_{rel}) – den man üblicher Weise als Relevanzschwelle bezeichnet (also Cohen’s d oder Hedge’s g im bisherigen IQWiG-Vorgehen) – und einem „kleinsten Effekt“ (δ_{min}), der den Einsatz einer Intervention noch rechtfertigen würde [6], weil Effekte größer (oder gleich) diesem δ_{min} bezüglich ihrer Relevanz für Patienten nicht ausgeschlossen werden können. Zur Illustration soll die Abbildung aus [5] dienen:



Auf individuellem Level haben diese beiden Schwellenwerte, δ_{rel} (oder d oder g) und δ_{min} ein Korrelat (keine direkte Entsprechung!), was wir nur erwähnen, um den Gedankengang klarer zu machen. δ_{rel} entspricht dabei einem Δ_{rel} auf individueller Ebene, welches dadurch definiert ist, dass das Vorenthalten einer Intervention mit einem Effekt von mindestens Δ_{rel} einen definitiven Nachteil für den Patienten darstellen würde (das entspricht IQWiG’s Wahl von d oder g = 0.2 auf Gruppenlevel).

δ_{min} entspricht einem Δ_{min}, einem kleinsten Effekt, ab dem man bei einem Patienten überhaupt von „Wirksamkeit“ sprechen kann. D.h. man kann nicht ausschließen, dass Effekte von mindestens δ_{min} für zumindest einige Patienten wahrnehmbar, also relevant sein

können. Damit ist also schon klar, dass δ_{min} der Schwellenwert sein sollte, gegen den man sich beim Testen mit verschobener Nullhypothese abgrenzen muss.
Wenn man hierzu noch den angenommenen Effekt für die Fallzahlplanung (Erwartungswert unter der Alternativhypothese), δ_{FZP}, einbezieht, ergibt sich folgende generelle Ungleichung:

$$\mathbf{0 \leq \delta_{min} \leq \delta_{rel} \leq \delta_{FZP}}$$

Wenn man nun die bisher und in diesem Berichtsplan vom IQWiG verwendete Relevanzschwelle von δ_{rel} = 0.2 einsetzt**,** ergibt sich sofort, dass ein δ_{min}, welches als shift-Parameter in der verschobenen Nullhypothese verwendet werden sollte, kleiner oder gleich diesem Wert 0.2 sein sollte und auch plausibel in diesem Kontext sein kann. Einer Absenkung von δ_{min} auf Werte < 0.2 ist also selbst dann nichts Prinzipielles entgegenzusetzen, wenn das IQWiG an seiner bereits 2009 selbst gewählten Forderung von δ_{rel} = 0.2 festhält. Dass die – prinzipiell mögliche - Wahl $\delta_{min} = \delta_{rel}$ bei gleichzeitiger Festsetzung von δ_{rel} = 0.2 hier i.d.R. nicht getroffen werden sollte, wird von mehreren Autoren deutlich gemacht [5, 6, 7, 8]. Dies beruht auf der inhaltlichen Überlegung zur Bedeutung dieser unterschiedlichen Schwellenwerte und der spezifischen Rolle von δ_{min} als dem Verschiebeparameter beim Testen mit einer um δ_{min} verschobenen Nullhypothese. Effekte, die kleiner als δ_{rel} sind, sind eben nicht per se irrelevant, erst Effekte die kleiner sind als δ_{min}, sind als irrelevant anzusehen.
Wie die Wahl von δ_{min} im Kontext der Nutzenbewertung im möglichen Intervall $0 \leq \delta_{min} < 0.2$ erfolgen sollte, diskutieren wir im nachfolgenden Abschnitt c).

c) $\delta_{min} = 0$ bei Nutzenbewertungen anhand patientenrelevanter Endpunkte

Zusätzlich zu den methodischen Ausführungen kann eine einfache Überlegung eine Anleitung zur Wahl von δ_{min} sein. In Nutzenbewertungen nach dem allgemeinen Methodenpapier des IQWiG und auch im konkreten Berichtsplan werden per se nur Endpunkte betrachtet, die patientenrelevant sind. Für diese Endpunkte wird ja schon immer gefordert, dass Effekte statistische Signifikanz aufweisen müssen, um zur Anerkenntnis eines Nutzenbelegs zu führen.

Alle beobachteten Effekte, die im Mittel und statistisch signifikant größer als 0 sind, beinhalten die Chance, zumindest für einen Anteil der Patienten eine relevante Verbesserung bereitzuhalten (da der Endpunkt selbst ja genau diesbezüglich ausgewählt

wurde, relevant zu sein), weshalb keine höhere Schwelle für die Anerkenntnis eines Nutzenbelegs in einem solchen Endpunkt gewählt werden sollte. Damit ist es u. E. klar, dass für alle patientenrelevanten Endpunkte, für die zusätzlich eine „Relevanzbetrachtung" vorgenommen werden soll, für die aber keine anerkannte „echte" (ankerbasierte) Relevanzschwelle existiert und die Hilfskonstruktion über eine SMD gegangen werden soll, ein $\delta_{min} = 0$ gelten sollte.

Diese Forderung wird weiter untermauert durch die Überlegung, wie im Rahmen des Prozesses aus Nutzenbewertung und Kosten-Nutzenbewertung beim IQWiG vorgegangen wird. Dies führen wir unter d) weiter aus.

d) Relevanz-Schwellenwerte $\delta_{min} > 0$ sind als zusätzliche Hürde vor dem Zugang zur Kosten-Nutzenbewertung unzulässig und nicht im Sinne einer effizienten Gesundheitsversorgung

Das Verfahren ist gegenwärtig so ausgestaltet, dass Interventionen, die in der Nutzenbewertung (mindestens) einen Nutzenbeleg erbracht haben, einer Kosten-Nutzen-Bewertung zugeführt werden können – andere jedoch nicht. D.h. nur Interventionen können ihre Kosteneffizienz nach den Verfahren der Kosten-Nutzen-Bewertung [IQWiG 2009] zeigen, wenn der Nutzenbeleg erfolgreich erbracht werden konnte.

Wenn nun aber durch die Wahl von $\delta_{min} > 0$ Nutzenbelege nicht anerkannt werden, welche aber trotzdem für einen Anteil der Patientenpopulation relevant sein können (so lange sie statistisch signifikant sind) und damit Interventionen gar nicht einer Kosten-Nutzenbewertung unterzogen würden, würde dies dazu führen, dass gewisse therapeutische Verbesserungen, die eventuell keine oder nur geringfügige Mehrkosten verursachen (oder sogar Kostenreduktionen beinhalten können), hierbei gar nicht bewertet würden und damit u.U. in der Erstattungsfrage noch nicht einmal die „Teilnahmehürde" überspringen können. Dies kann nicht im Interesse des Gesundheitssystems und insbesondere der Patienten sein. Die Anwendung von frühzeitig (im Rahmen der Nutzenbewertung) erhöhten Relevanzschwellen ist also ganz eindeutig ein Eingreifen, welches nicht ergebnisoffen die Erstattbarkeit von Interventionen einschränkt und damit mögliche (kleinere) therapeutische Fortschritte, die sich darüber hinaus als kosteneffizient erweisen können, blockiert. Dies ist ein weiteres gewichtiges Argument für die schon weiter oben begründete Wahl von $\delta_{min} = 0$ bei Relevanzbeurteilungen über SMDs in den unter c) eingegrenzten Situationen und Endpunkten.

8. Genauere Spezifikationen zu Kriterien unter denen Endpunkte nicht einbezogen werden, Unterschiede der relativen Häufigkeit fehlender Werte zwischen Vergleichsgruppen

In 4.4.1 wird ausgeführt, dass für die Nutzenbewertung einzelne patientenrelevante Endpunkte nicht berücksichtigt werden, „insbesondere“ wenn „viele Patienten nicht in der Auswertung enthalten sind“. Weiter wird spezifiziert, dass dies „in der Regel“ dann eintritt, wenn diese Auswertungen auf „weniger als 70% der in die Auswertung einzuschließenden Patienten basieren“.

Wir bitten darum, im Berichtsplan klar zu definieren, welche Gründe außer der Nichtverfügbarkeit von Patientendaten mit einer Rate von mehr als 30% zur Nichtberücksichtigung von Endpunkten führen können (Erklärung des „insbesondere“).

Weiterhin sollte für diese o.g. 70%-Regel ebenso wie für die „15 Prozentpunkt-Regel“ weiter unten in 4.4.1 (maximal zulässige Rate inkrementeller Drop-out-Fälle zwischen den Behandlungsgruppen) angegeben werden, welches die Basis (der Nenner) bei der Berechnung dieser Kriterien darstellt. Soll der Stichprobenumfang der ITT-Population der Gesamtstudie oder der jeweiligen Behandlungsgruppe in der Studie benutzt werden oder werden andere Möglichkeiten für die Definition des jeweiligen Nenners benutzt?

9. Effektmaße bei Meta-Analysen

In 4.4 und 4.4.2 wird auf die Effektmaße eingegangen, die bei Meta-Analysen Verwendung finden sollen. Für kontinuierliche Variablen soll „gegebenenfalls“ eine standardisierte Mittelwertdifferenz (SMD) eingesetzt werden. Es soll Hedges‘ g dazu herangezogen werden.

Wir bitten um Klarstellung, über die Rolle der Verwendung einer SMD in Meta-Analysen einer kontinuierlichen Variablen. Wenn eine SMD-Betrachtung vorgenommen wird, sollte diese zusätzlich zur Betrachtung der Originalskala der kontinuierlichen Variablen erfolgen und nicht alternativ zu dieser.

Darüber hinaus sollte geklärt werden, welche Variante von Hedge’s g bei dieser Nutzenbewertung eingesetzt werden soll, da verschiedene Varianten der Adjustierung von Cohen’s d existieren.

Ebenfalls sollte im Berichtsplan festgelegt werden, in welchen Situationen Effektschätzer aus Einzelstudien und insbesondere deren Standardabweichungen in einer Meta-Analyse mittels solcher SMDs gemeinsam ausgewertet werden können und sollen und wie beim Vorliegen unterschiedlicher Situationen in Einzelstudien vorgegangen werden soll (Mittelwertvergleich von prä-post Veränderungen – adjustiert oder nicht-adjustiert für Covariable - ; Mittelwertvergleiche von Endpunkten (d.h. keine prä-post Veränderungen) – adjustiert oder nicht-adjustiert für Covariable).

Literatur

(1) Guyatt G, Cook D, Jaeschke R, Schünemann H, Pauker S. Grading recommendations: a qualitative approach. In: Guyatt G, Rennie D, editors. Users' guides to the medical literature: a manual for evidence based practice. Chicago, IL: AMA Press; 2002. 599–608.

(2) Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidem. 2008; 61: 102-109

(3) D'Agostino, RB Snr. Editorial: Quantifying the comparison of two groups. Stats in Med. 1999; 18: 2551-2555

(4) Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, Glasziou P, Guyatt GH. Interpreting results and drawing conclusions. In: Higgins JP, Green S (Ed). Cochrane handbook for systematic reviews of interventions. Chichester: Wiley; 2008. 359-387.

(5) Thomas S. Klinische Relevanz von Therapieeffekten: Systematische Sichtung, Klassifizierung und Bewertung methodischer Konzepte. Diss. Universität Duisburg-Essen, 2009

(6) Victor N. On clinically relevant differences and shifted nullhypotheses. Meth Inform Med 1987; 26: 109-116

(7) Braitman LE. Confidence intervals assess both clinical significance and statistical significance. Ann Int Med 1991: 114; 515-517

(8) Kieser M, Hauschke D. Assessment of clinical relevance by considering point estimates and associated confidence intervals. Pharmaceut Statist. 2005; 4: 101-107;

A 1.4 Novartis Pharma GmbH

Autoren:
Färber, Lothar, PD, Dr.
Krippner, Felix
Wasmuth, Timo
Peters, Lars

Adresse:
Novartis Pharma GmbH
Roonstrasse 25
90429 Nürnberg

Stellungnahme der Novartis Pharma GmbH zur Auftragsnummer A05-18

„Tiotropiumbromid bei COPD“

Anmerkungen zur projektspezifischen Methodik unter jeweiliger Angabe wissenschaftlicher Literatur zur Begründung der Anmerkung

1. Endpunkte

In Abschnitt 4.1.3 des Berichtsplans werden Endpunkte genannt, die bei der Untersuchung berücksichtigt werden sollen. Es wurde jedoch verzichtet, Instrumente zu definieren, die zur Messung der Endpunkte zugelassen sind. Daher möchten wir anregen, die folgenden Messinstrumente explizit im Berichtsplan einzuschließen:

1.1. TDI als Messinstrument für Dyspnoe

Dyspnoe oder Atemlosigkeit ist ein wichtiges limitierendes Symptom, unter dem COPD-Patienten leiden. Eins der etabliertesten Instrumente zur Messung der Atemnot ist der Transitional Dyspnoea Index (TDI) [Jones, 2005]. Hierbei handelt es sich um ein multidimensionale Instrument, das die Differenz zum Baseline Dyspnoe Index (BDI) misst. Der BDI beinhaltet drei Komponenten (Veränderungen der funktionellen Beeinträchtigung, Veränderungen der Belastbarkeit, Veränderungen der Größe der Anstrengung), die auf einer Skala von 0 (schwere Dyspnoe) bis 4 (keine Dyspnoe) bewertet werden. Der TDI ändert sich in jeder Komponente des BDI im Vergleich zur Baseline auf einer Skala von +3 (große Verbesserung) bis zu -3 (große Verschlechterung). Es konnte gezeigt werden, dass der TDI ein valides, verlässliches und sensitives Instrument ist. Der Schwellenwert für eine klinisch relevante Verbesserung (MCID) liegt bei ≥1 Einheit [Mahler, 1984; Mahler, 2006; Witek, 2003].

Daher sollte der TDI als Messinstrument der Dyspnoe als COPD-Symptom explizit eingeschlossen werden.

1.2. Verbrauch an Bedarfsmedikation

Der Verbrauch an Bedarfsmedikation (Short Acting Beta-2 Agonists [SABAs] oder Short Acting Muscarinic Antagonist [SAMA]) reflektiert den Grad und/oder die Häufigkeit von auftretenden Symptomen für den Patienten und demnach auf die Fähigkeit der Intervention die COPD-spezifischen Symptomatik zu kontrollieren.

Der Bedarf kann über die Gesamtzahl bzw. am Anteil der Tage gemessen werden, an denen keine Bedarfsmedikation verwendet wurde, oder anhand der durchschnittlichen Anzahl der verbrauchten Hübe pro Tag.

Hierbei muss jedoch beachtet werden, dass es keine vordefinierten Schwellenwerte gibt, ab denen eine klinisch relevante Verbesserung vorliegt. Vielmehr sollte jegliche statistische

signifikante Reduktion des Verbrauchs an Bedarfsmedikation als erstrebenswert in der COPD-Therapie angesehen werden.

Daher ist der Verbrauch an Bedarfsmedikation ein weiterer Parameter, der bei der Bewertung Berücksichtigung finden sollte.

1.3. SGRQ als Messinstrument für gesundheitsbezogene Lebensqualität

Der St. George's Respiratory Questionnaire (SGRQ) ist ein validiertes Instrument zur Messung der gesundheitsbezogenen Lebensqualität, der von den COPD-Patienten selbst ausgefüllt wird [Jones, 1992] und der am häufigsten verwendete Fragebogen zur Erfassung der Lebensqualität [Gillissen, 2008].

Der SGRQ misst die Veränderung der Lebensqualität zwischen zwei Zeitpunkten. Er besteht aus 50 Fragen, aufgeteilt in drei Komponenten: Symptome, Aktivität und Beeinträchtigung, die in einem Gesamtscore auf einer Skala von 0 (am besten) bis 100 (am schlechtesten) zusammengefasst werden. Als MCID für dieses Messinstrument gilt eine Verbesserung von vier Einheiten im Vergleich zu Plazebo oder zur Baseline [Jones, 2005]. Eine solche Verbesserung zeigt sich typischerweise in folgenden Formen:

1) Der Patient benötigt nicht mehr zusätzliche Zeit zum sich waschen oder anziehen, kann die Treppen steigen ohne anzuhalten und zur Unterhaltung ausgehen
2) Der Patient empfindet, dass viele Dinge nicht mehr besonders viel Mühe brauchen; er muss bei Hausarbeit keine Pausen mehr einlegen und kann Gegenstände die Treppe hochtragen
3) Der Patient muss nicht mehr langsamer gehen als andere und ist nicht mehr außer Atem beim Waschen, Anziehen oder beim Überbeugen [Jones, 2002].

Der SGRQ eignet sich also sehr gut zur Messung der gesundheitsbezogenen Lebensqualität bei COPD-Patienten und sollte daher explizit als anerkannter Parameter in der Bewertung berücksichtigt werden. Anzumerken bleibt, dass in der jüngsten Vergangenheit die MCID von vier Einheiten kritisch hinterfragt wird und die Tendenz zu einem geringeren Schwellenwert abzusehen ist.

1.4. körperliche Belastbarkeit

Die bei der COPD zur Untersuchung körperlicher Belastbarkeit eingesetzten Methoden sind im Wesentlichen der 6-Minute Walking Test (6MWT) sowie die Kardiopulmonale Belastungstestung (Cardiopulmonary Exercise Testing [CPET]).

Der 6MWT ist ein relativ einfach durchzuführender Test, bei dem die Gehstrecke erfasst wird, die ein Patient in 6 min in einer selbstbestimmten Geschwindigkeit gehen kann [Salzman, 2009]. Die standardisierte Durchführung ist in einem internationalen Positionspapier der American Thoracic Society festgelegt [ATS Statement, 2002]. Mittels des 6MWT soll die Fähigkeit der Patienten abgebildet werden, alltägliche Aufgaben zu erfüllen. Durch die Veränderung der Gehstrecke kann in diesem Test der Erfolg von medizinischen Interventionen bei Patienten mit moderaten bis schweren Lungenerkrankungen gemessen werden [ATS Statement, 2002]. Außerdem wird in 6MWT ein guter Prädiktor für Mortalität

gesehen [Casanova, 2008]. Umstritten sind allerdings die Anforderungen an die Veränderung der Wegstrecke, um von einer klinischen Relevanz sprechen zu können (= minimal clinical important difference = MCID). Für Patienten mit COPD werden Werte um 35 Meter als klinisch relevant diskutiert [Salzman, 2009].

Als weitere wissenschaftliche Methode zur Erfassung körperlicher Belastbarkeit bei Patienten mit COPD ist die CPET (Cardiopulmonary Exercise Testing) anerkannt. Bei diesem Verfahren werden die Patienten mittels eines Fahrradergometers oder einer Tretmühle körperlicher Belastung unterschiedlicher Intensität ausgesetzt. Es kann neben der maximalen Leistung und Ausdauer der Patienten auch der Gasaustausch in der Lunge, kardiale Parameter untersucht sowie eine Blutgasanalyse durchgeführt werden. Methodik und erfassbare Parameter sind im gemeinsamen Positionspapier der American Thoracic Society und American College of Chest Physicians (=ATS/ACCP) festgelegt [ATS/ACCP, 2003]. Aber auch bei der CPET fehlt ein einheitliches Verständnis für klinisch relevante Veränderung in den Parametern [ATS/ACCP, 2003]. Untersuchungen hinsichtlich Ausdauer unter konstanter Belastung lassen eine MCID in einer 30%igen Steigerung der Belastungsdauer vermuten [Puente-Maestru, 2009].

Aufgrund der unterschiedlichen Möglichkeiten zur Erfassung von körperlicher Belastbarkeit ist es notwendig, dass im Berichtsplan hinsichtlich Akzeptanz der Methoden und Parameter Stellung genommen wird, sowie Angaben zu den geforderten Schwellenwerten vorgenommen werden.

1.5. Lungenfunktion als Prädikator für patientenrelevante Endpunkte

Spirometrisch erfasste Parameter sind die Basis der Diagnostik, Schweregradeinteilung und Verlaufskontrolle bei der COPD und anderen pneumologischen Erkrankungen. Das forcierte expiratorische Volumen innerhalb einer Sekunde (FEV_1) sowie die inspiratorische Kapazität (IC) sind wichtige Parameter in der Beurteilung des Krankheitszustands eines COPD Patienten.

FEV_1

Der FEV_1-Wert ist als der zentrale diagnostische Parameter zur Untersuchung der COPD allgemein anerkannt. So baut die Differenzierung von Asthma und COPD weitestgehend auf dem Tiffenau-Index (FEV_1/FVC) auf. Auch die im Berichtsplan zitierte Schweregradeinteilung entsprechend der Leitlinie der *Global Initiative for Chronic Obstructive Pulmonary Disease* (GOLD) basiert auf dem FEV_1 des Patienten im Verhältnis zum Normwert [GOLD, 2009].

Doch auch über die Schweregradeinteilung hinaus ist der FEV_1 der zentrale Parameter bei Diagnose und Verlaufskontrolle der COPD. Die internationalen GOLD-Leitlinien würdigen die Spirometrie und damit den FEV_1 als „… the gold standard as it is the most reproducible, standardized and objective way of measuring airflow limitation." [„... den Goldstandard, da es ein reproduzierbares, standardisiertes und objektives Instrument zur Messung der Atmungsbeeinträchtigung ist."] [GOLD, 2009]

Die Durchführung der spirometrischen Erfassung des FEV_1 ist im höchsten Maße standardisiert, wodurch eine Vergleichbarkeit auch zwischen unterschiedlichen Studien gewährleistet ist [Miller, 2005].

Nicht zuletzt deshalb fordert die EMA den FEV_1 als primären Endpunkt in Studien, die zur Zulassung von Arzneimitteln zur Behandlung der COPD führen sollen [CPMP, 1999], weshalb er als der etablierte Standard zur Feststellung der Wirksamkeit von Arzneimitteln anzusehen ist.

Da ein erniedrigter FEV_1-Wert mit einer erhöhten Gesamtletalität assoziiert wird [Gillissen, 2005], sollte dieser Parameter in der Bewertung Berücksichtigung finden.

Inspiratorische Kapazität (IC)

Die Beeinträchtigung der Lungenfunktion bei COPD führt zur Überblähung der Lunge. Dies erhöht die inspiratorische Ladung und induziert funktionale Schwäche des Zwerchfells und der Atemwegsmuskulatur. Bei Anstrengung verursacht dies schnelles und flaches Atmen und eine fortschreitende Reduktion der dynamischen Lungenfüllung, was eine weitere Verringerung der Belastungstoleranz, sowie Belüftungsstörung zur Folge hat [O'Donnell, 2006]. Es wurde gezeigt, dass die Überblähung der Lunge ein unabhängiger Prädikator für Mortalität bei COPD ist (Atemwegsinduziert und all-cause) [Casanova et al., 2005]. Wie der FEV_1 ist auch die Messung der IC klar definiert, wodurch auch bei diesem Parameter ein hoher Standard gewährleistet ist [Miller, 2005].

Wir plädieren daher für die Aufnahme von FEV_1 und IC als prädiktive Parameter für patientenrelevante Endpunkte in die Liste der untersuchten Endpunkte.

2. *Studiendauer*

Im vorläufigen Berichtsplan wird als Einschlusskriterium eine Studiendauer von mindestens sechs Monaten unter Berufung auf die CPMP-Leitlinien genannt. Die vom CPMP gemachte Angabe bezieht sich jedoch auf die Messung von COPD-Symptomen und kann daher nicht für alle Parameter verwendet werden.

Ein weiteres Dokument des Committee for Medicinal Products for Human Use (CHMP) zur Messung von gesundheitsbezogener Lebensqualität empfiehlt eine Studienlänge von 3 - 6 Monate, was bestätigt, dass für diesen Parameter bereits kürzere Studien valide Messergebnisse liefern können [CHMP, 2005].

Insbesondere bei den weiteren Parametern FEV_1 und IC können bereits kürzere Studien deutliche Effekte zeigen.

Daher sollten das entsprechende Einschlusskriterium geändert werden, sodass auch 3-monatige Studien für die entsprechenden Parameter berücksichtigt werden.

Verweise auf qualitativ angemessene Unterlagen, insbesondere bislang unpublizierte Daten, einschließlich einer Begründung für ihre jeweilige fragestellungsbezogene Eignung und Validität

Zum gegenwärtigen Zeitpunkt gibt es hierzu keine Kommentare.

Literaturangaben

American Thoracic Society. ATS statement: guidelines for the six-minute walk test. Am J Respir Crit Care Med 2002; 166:111–117

ATS/ACCP Statement on Cardiopulmonary Exercise Testing. Am J Respir Crit Care Med 2003; 167: 211–277

Casanova C, Cote C, Marin JM, et al. Distance and oxygen desaturation during the 6-min walk test as predictors of long-term mortality in patients with COPD. Chest 2008; 134:746–752

Casanova C, Cote C, de Torres JP et al. Inspiratory-to-total lung capacity ratio predicts mortality in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2005;171:591-97.

CHMP. Committee for Medicinal Products for Human Use. Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products. 2005

CPMP. Committee for Proprietary Medicinal Products. Points to consider in on clinical investigations of medicinal products in the chronic treatment of patients with chronic obstructive pulmonary disease (COPD). 1999.

Gillissen A, Buhl R, Kardos P, Puhan M, Rabe KF, Rothe T, Sauer R, Welte T, Worth H, Menz G. Studienendpunkte bei der chronisch-obstruktiven Lungenerkrankung (COPD): „Minimal Clinically Important Difference“. *Pneumologie* 2008;62:149-57

GOLD. Global strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease. 2009.

Jones PW, Lareau S, Mahler DA. Measuring the effects of COPD on the patient. *Respir Med* 2005;99 Suppl B:11-S18.

Jones PW. Interpreting thresholds for a clinically significant change in health status in asthma and COPD. *Eur Respir J* 2002;19:398-404.

Jones PW. St. George's Respiratory Questionnaire: MCID. *COPD* 2005;2:75-79.

Jones PW, Quirk FH, Baveystock CM, Littlejohns P. A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire. *Am Rev Respir Dis* 1992;145:1321-27.

Mahler DA. Mechanisms and measurement of dyspnea in chronic obstructive pulmonary disease. *Proc Am Thorac Soc* 2006;3:234-38.

Mahler DA, Weinberg DH, Wells CK, Feinstein AR. The measurement of dyspnea. Contents, interobserver agreement, and physiologic correlates of two new clinical indexes. *Chest* 1984;85:751-58.

Miller MR et al. ATS/ERS Standardization of Lung Function Testing: Standardization of Spirometry. Eur Respir J 2005;26:319-338

O'Donnell DE. Hyperinflation, dyspnea, and exercise intolerance in chronic obstructive pulmonary disease. *Proc Am Thorac Soc* 2006;3:180-184.

Puente-Maestu L, Villar F, de Miguel J, Stringer WW, Sanz P, Sanz ML, Garcia de Pedro J, Martinez-Abad Y. Clinical relevance of constant power exercise duration changes in COPD. Eur Respir J 2009;34:340–345

Salzman SH. The 6-Min Walk Test. Chest 2009;135:1345-1352

Witek TJ, Jr., Mahler DA. Minimal important difference of the transition dyspnoea index in a multinational clinical trial. *Eur Respir J* 2003;21:267-72.

A 1.5 Pfizer Deutschland GmbH

Autoren:
Hohmann, Christoph, Dr.
Leverkus, Friedhelm
Wernitz, Martin, Dr.

Adresse:
Pfizer Deutschland GmbH
Linkstraße 10
10785 Berlin

Stellungnahme der Firma Pfizer Pharma GmbH (im Folgenden „Pfizer) zum vorläufigen Berichtsplan der Bewertung A05-18 „Tiotropiumbromid bei COPD“

Einleitung

Das IQWiG hat am 19.04.2010 den vorläufigen Berichtsplan zum Auftrag A05-18 „Tiotropiumbromid bei COPD“ veröffentlicht.

Pfizer kommentiert hiermit den vorläufigen Berichtsplan.

Wir begrüßen ausdrücklich die Einbindung von Patientenorganisationen in den vorliegenden IQWiG-Prozess. Angesichts der grundlegenden Unterschiede zwischen der COPD und dem Asthma bronchiale ist jedoch nicht nachvollziehbar, warum während der Erstellung des Berichtsplans als Patientenvertreter Mitglieder des Deutschen Allergie- und Asthmabundes e.V. einbezogen wurden, auf dessen Internetseiten unter http://www.daab.de/index.php keinerlei Hinweis darauf zu finden ist, dass der DAAB eine Interessenvertretung von COPD-Patienten darstellt. Patientenverbände, die sich explizit den Belangen von COPD-Patienten verschrieben haben, wie beispielsweise die Selbsthilfegruppe Lungenemphysem COPD Deutschland (http://lungenemphysem-copd.de/), oder die COPD Selbsthilfe e.V. (http://copd-selbsthilfe.de) wurden hingegen nicht konsultiert.

1 Präzisierung der Zielgröße "gesundheitsbezogene Lebensqualität"

Im vorläufigen Berichtsplan wird als eine der Zielgrößen "gesundheitsbezogene Lebensqualität" genannt. Dieser patientenberichtete Endpunkt hat zweifelsohne große Relevanz für die Bewertung des Nutzens einer diagnostischen oder therapeutischen Maßnahme. Aus der knappen Benennung dieses Endpunktes wird jedoch nicht ausreichend klar, wie das IQWiG diese Zielgröße in der Nutzenbewertung genau auffassen wird. In mehreren früheren Nutzenbewertungen hat das IQWiG diesen Punkt präziser gefasst. Als Beispiele sind nachfolgend einige solcher Bewertungsaufträge angeführt:

Bewertungsauftrag	**Festlegung der Zielgröße**
A05-04 (Kurzwirksame Insulinanaloga zur Behandlung des Diabetes mellitus Typ 2)	Erhalt bzw. Besserung krankheitsbezogener Lebensqualität
A05-13 (Fixe Kombinationen aus Kortikosteroiden und lang wirksamen Beta-2 Rezeptoragonisten zur inhalativen Anwendung bei Patienten mit Asthma bronchiale)	Besserung bzw. Erhalt der gesundheitsbezogenen Lebensqualität
A05-14 (Leukotrien-Rezeptor-Antagonisten bei Patienten mit Asthma bronchiale)	Besserung bzw. Erhalt der erkrankungsbezogenen Lebensqualität
A05-19A (Cholinesterasehemmer bei Alzheimer Demenz)	Besserung bzw. Erhalt der krankheitsbezogenen Lebensqualität; Besserung bzw. Erhalt der Lebensqualität der (betreuenden) Angehörigen

Bei einer chronisch progredienten Erkrankung wie der COPD nimmt die Lebensqualität im Verlauf von Monaten und Jahren im Mittel ab. Unter diesen Umständen ist auch von einer effektiven Therapie eine „Verbesserung" der Lebensqualität (gegenüber dem Ausgangszustand bei Studienbeginn) nicht unbedingt zu erwarten. Je länger die Studiendauer ist, desto ausgeprägter wird der klinische Effekt einer Therapiemaßnahme von diesem nachteiligen Spontanverlauf der Erkrankung überlagert, sodass sich mit längerer Studiendauer das Therapieziel zunehmend von der „Verbesserung" zum „Erhalt" der Lebensqualität verlagert.

Daher bitten wir um eine Präzisierung der Zielgröße in der Weise, wie es auch in den o.g. anderen Bewertungsaufträgen erfolgt ist. Desweiteren sollte der Begriff „Lebensqualität“ durch den des „Gesundheitsstatus“ ergänzt werden, da die auf dem Gebiet der COPD gebräuchlichen Fragebogeninstrumente (wie z.B. der St. George’s Respiratory Questionnaire – SGRQ) genau genommen ein quantitatives Maß für den Gesundheitsstatus und nicht die Lebensqualität ermitteln[6].

Die Zielgröße „gesundheitsbezogene Lebensqualität“ sollte demnach folgendermaßen formuliert sein: „Verbesserung bzw. Erhalt der gesundheitsbezogenen Lebensqualität / des Gesundheitsstatus“

2 Gegenüberstellung der Ergebnisse der Einzelstudien: Nichtberücksichtigung von Studienergebnissen, die auf weniger als 70% der randomisierten Patienten beruhen

Unter Punkt 4.4.1 des vorläufigen Berichtsplans wird ausgeführt, dass Ergebnisse i. d. R. nicht in die Nutzenbewertung einfließen, „wenn diese auf weniger als 70% der in die Auswertung einzuschließenden Patienten basieren, d. h. wenn der Anteil der fehlenden Werte größer als 30% ist“. Diesbezüglich wird Bezug genommen auf die Arbeit von Schulz und Grimes[7], die einen Prozentsatz von 20 als oberste Grenze für einen noch tolerablen „loss to follow-up“ hinsichtlich der Validität einer Studie zitieren.

Insbesondere bitten wir um Klarstellung, dass sich die genannten Zahlen nicht auf die vorzeitigen Abbruchraten beziehen. Diese sind von der Länge der Studie abhängig und würden Langzeitstudien fast zwangsläufig invalide machen. Wir gehen vielmehr davon aus, dass sich die Zahlen auf den Anteil an Patienten beziehen, die nicht in die Auswertung eingeschlossen werden. Diesbezüglich bitten wir um Konkretisierung bzw. Präzisierung wie dies exakt zu verstehen ist, analog zur Anhörung des vorläufigen Berichtsplan zum Auftrag A09-01 (Dipyridamol + ASS zur Sekundärprävention nach Schlaganfall oder TIA). Das Problem von ggf. fehlenden Werten muss mit geeigneten statistischen Methoden adressiert werden.

Die Vorgabe im Vorläufigen Berichtsplan zum Bewertungsauftrag A05-18, dass Ergebnisse i. d. R. nicht in die Nutzenbewertung einfließen, „wenn diese auf weniger als 70% der in die Auswertung einzuschließenden Patienten basieren“, sollte dahingehend konkretisiert werden, dass sich dies auf den Anteil der Patienten, die in eine Analyse eingeschlossen werden, bezieht.

3 Informationssynthese und –analyse

Das IQWiG sieht im vorläufigen Berichtsplan die Bewertung der klinischen Relevanz eines beobachteten Patientennutzens vor. Dabei soll die Relevanzbewertung je nach Verfügbarkeit der Daten auf Basis von Mittelwertsdifferenzen und / oder anhand von Responderanalysen unter primärer Berücksichtigung von validierten MIDs (minimal important differences) erfolgen.

Hierbei wird jedoch nicht mitgeteilt, welche Relevanzgrenzen es pro Endpunkt gibt – und ob diese wohlbegründet sind. Dies gilt insbesondere für die Erfassung der Zielgrößen „COPD-Symptome“ und „gesundheitsbezogene Lebensqualität“. Bei COPD-Symptomen und bei der gesundheitsbezogenen Lebensqualität handelt es sich um subjektive Endpunkte, die in klinischen Studien in der Regel in Form von Fragebogen-basierten Skalen oder Scoring-Systemen erfasst werden. Zur Frage der Validität solcher Skalen oder Scoring-Systeme bzw. zum Vorliegen entsprechender Relevanzschwellen (minimal important differences – MIDs) hätte im Rahmen eines Scoping-Prozesses im Vorfeld der Berichtsplanerstellung eine Befragung von klinisch ausgewiesenen Experten der wissenschaftlich-medizinischen Fachgesellschaften im für die Fragestellung relevanten Therapiegebiet erfolgen sollen, wie dies in nationalen und internationalen Gutachten[1,2,6] wiederholt gefordert wurde.

Liegen für (komplexe) Skalen keine validierten Grenzen vor, so bezieht sich das IQWiG auf Hedges’ g von 0.2. Hedges’ g ist eine Modifikation von Cohen’s d. Es ist ein rein statistisches Maß, das zur Vereinfachung der Fallzahlabschätzung beschrieben wurde. Cohen bemerkt, dass mit der Einführung der Konvention, Effekte in klein, mittel und groß zu kategorisieren, nicht die Wichtigkeit (klinische Relevanz) von kleinen, mittleren oder großen Effekten präjudiziert ist. Er warnt vor einer Missdeutung und betont die Relativität der Kategorien in verschiedenen Anwendungsfeldern[3].

Die europäische Arzneimittelbehörde EMA lehnt bei der Bestimmung eines „non-inferiority margins“ solche Effektgrößen als alleinige Grundlage ab und fordert stattdessen eine Interpretation im klinischen Kontext: „It is not appropriate to use effect size (treatment difference divided by standard deviation) as justification for the choice of non-inferiority margin. This statistic provides information on how difficult a difference would be to detect, but does not help justify the clinical relevance of the difference, and does not ensure that the test product is superior to placebo.“[4]

Das IQWiG will – falls Responderauswertungen nicht zur Verfügung stehen – einen Test mit verschobenen Nullhypothesen anwenden, indem es fordert, dass das zum beobachteten Effekt korrespondierende Konfidenzintervall vollständig oberhalb der Relevanzschwelle liegt, damit von einer relevanten Effektstärke ausgegangen werden kann. Dieser Test ist bei klinischen Prüfungen, außer im Nicht-Unterlegenheits-Design, unüblich. Bei dieser Vorgehensweise ist der Nachweis, dass eine MID statistisch signifikant überschritten wird, nur mit sehr großen Fallzahlen zu führen. Die MID, die durchaus relevant ist, kann praktisch nicht nachgewiesen werden.

Wir halten die Einbeziehung von Fachgesellschaften im Rahmen eines Scoping-Prozesses zur Frage der Beurteilung der klinischen Relevanz eines Patientennutzens für sehr wichtig. Der Wert der Effektgröße (Hedges' g) von 0,2, den das IQWiG als Minimalwert einer klinischen Relevanz für alle Messskalen und Patientenpopulationen ansieht, stellt aus unserer Sicht keinen international verbindlichen Richtwert dar. Vielmehr ist dieser Ansatz rein technischer Natur und soll mit einem statistischen Test eine binäre Entscheidung treffen. Letztlich ist aber jede statistisch signifikante Verbesserung eines patientenrelevanten Endpunkts als Zusatznutzen aus klinischer Perspektive positiv zu bewerten.

4 Literaturverzeichnis:

1. Antes G, Jöckel KH, Kohlmann T, Raspe H, Wasem J. Kommentierende Synopse der Fachpositionen zur Kosten-Nutzenbewertung von Arzneimitteln – Erstellt im Auftrag des Bundesministeriums für Gesundheit 2007
2. Bekkering GE, Kleijnen J. Procedures and methods of benefit assessments for medicines in Germany. Eur J Health Econ 2008; 9 (Suppl 1): 5-29 bzw. Dtsch Med Wochenschr 2008; 133 (Suppl 7): S225-S246
3. Cohen J. Statistical power analysis for the behavioral sciences, 2nd Edition. Taylor & Francis Group, New York 1988
4. EMA – Committee for medicinal products for human use (CHMP). Guideline on the choice of inferiority margin 2005
5. Jones PW. Health status and the spiral of decline. COPD 2009; 6: 59-63
6. NICE – National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal. http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf (letzter Zugriff: 17.05.2010)
7. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet 2002; 359: 781-785