IQWiG Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Institute for Quality and Efficiency in Health Care

Dokumentation und Würdigung der Stellungnahmen zur „Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“

Version 1.0 vom 28.11.2013

Anschrift des Herausgebers:

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Im Mediapark 8 (KölnTurm)
50670 Köln

Tel.: +49 (0)221 – 35685-0
Fax: +49 (0)221 – 35685-1
E-Mail: methoden@iqwig.de
Internet: www.iqwig.de

Inhaltsverzeichnis

Seite

1 Dokumentation der Anhörung

Am 18.04.2013 wurde das Dokument zur Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie zu neuen Abschnitten zur Erstellung der Allgemeinen Methoden Version 4.1 veröffentlicht und zur Anhörung gestellt. Bis zum 22.05.2013 konnten schriftliche Stellungnahmen eingereicht werden. Insgesamt wurden 27 Stellungnahmen form- und fristgerecht abgegeben. Diese Stellungnahmen sind im Anhang abgebildet.

Unklare Aspekte in den schriftlichen Stellungnahmen wurden in einer wissenschaftlichen Erörterung am 30.09.2013 im Institut diskutiert. Das Wortprotokoll der Erörterung befindet sich in Kapitel 4.

2 Würdigung der Anhörung

Im Stellungnahmeverfahren zum Dokument „Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“ vom 18.04.2013 gingen bis zum 23.05.2013 Stellungnahmen von 27 Organisationen, Institutionen, Firmen und Einzelpersonen ein. Alle Stellungnahmen wurden in der Überarbeitung des Methodenpapiers berücksichtigt. In diesem Dokument werden ausschließlich die Kommentare aus den Stellungnahmen gewürdigt, die sich auf die aktualisierten bzw. neuen Abschnitte der Allgemeinen Methoden beziehen. Sofern sich aus den Stellungnahmen Änderungen in diesen Abschnitten ergaben, ist dies im vorliegenden Dokument explizit beschrieben.

2.1 Würdigung allgemeiner Stellungnahmen

2.1.1 Übertragbarkeit

Ein Stellungnehmender adressiert das Problem der Übertragbarkeit und schlägt vor, Verfahren zur Interpolation zuzulassen, ohne diese allerdings konkret zu benennen.

Grundsätzlich entstehen durch zahlreiche Mechanismen Einschränkungen oder auch Erweiterungen von Fragestellungen, die in den der entsprechenden Bewertung zugrunde liegenden (Primär-)Studien nicht berücksichtigt wurden oder auch nicht berücksichtigt werden konnten, z. B. die Einschränkung der Anwendung von Arzneimitteln aufgrund der Zulassung. Und es stellt sich in diesem Zusammenhang die Frage, ob nicht auch Daten zur Bewertung herangezogen werden können, die nicht exakt der eigentlich zu bewertenden Fragestellung entsprechen, z. B. im Hinblick auf die Patientenpopulation oder die Anwendung der Prüf- und / oder Vergleichsintervention. Im Methodenpapier des IQWiG und in vorausgegangenen Bewertungen werden 3 Vorgehensweisen benannt, wie mit dem Problem umgegangen werden kann. Der sprachlichen Einfachheit halber wird dies exemplarisch für die Frage der Übertragbarkeit von Daten von nicht zulassungskonform behandelten Patienten (Nichtzulassungspopulation) dargestellt, es lässt sich aber im Grundsatz verallgemeinern.

Die Studienpopulation wird herangezogen, wenn die Nichtzulassungspopulation innerhalb der Studienpopulation genügend klein ist, d. h. einen Anteil von weniger als 20 % ausmacht.

Die Zulassungspopulation innerhalb der Studienpopulation wird als Subgruppe betrachtet und gesondert ausgewertet. Wie vom Stellungnehmenden bemerkt, führt dies gegebenenfalls zu einem Powerverlust und zu einer Verminderung der Präzision. Das ist aber auch folgerichtig, da die Datengrundlage für die eigentliche Fragestellung reduziert ist und somit eine größere Unsicherheit resultiert. Um dem entgegenzuwirken, bleibt aber noch eine weitere (zusätzliche) Vorgehensweise:

Es ist durch geeignete Daten nachzuweisen, dass Ergebnisse aus der Nichtzulassungspopulation dennoch anwendbar sind. Hier bieten sich (statistische) Interaktionstests und der Ausschluss einer zumindest qualitativen Interaktion an (gegebenenfalls auch der Ausschluss

einer an den Schwellen für bestimmte Ausmaße orientierten quantitativen Interaktion bestimmter Größenordnung). Dies ist allerdings für alle relevanten Endpunkte nachzuweisen und beinhaltet auch die Darstellung von Ergebnissen (Schätzer und Präzisionsmaß) innerhalb der Subgruppen der Zulassungs- und Nichtzulassungspopulation. Kann eine relevante Interaktion ausgeschlossen werden, können auch die Ergebnisse aus der Nicht-zulassungspopulation für die Bewertung herangezogen werden. Letztlich bedeutet diese letzte Vorgehensweise eine Operationalisierung der vom Stellungnehmenden vorgeschlagenen „Interpolation".

Zusammenfassend ergibt sich, dass die beschriebenen Vorgehensweisen bereits im Methoden-papier enthalten sind.

2.1.2 Festlegungen des Instituts

Das Institut trifft an verschiedenen Stellen konkrete Festlegungen, z. B. bei der Definition, in welchen Situationen geschätzte Effekte gleichgerichtet sind und wann nicht. Diese Festlegungen werden in mehreren Stellungnahmen kritisiert, weil eine entsprechende Rationale fehlt, bzw. es wird gefordert, die Richtigkeit solcher Festlegungen zunächst durch empirische Resultate sicherzustellen.

Diese Kritik steht im Widerspruch zu anderen Stellungnahmen, in denen konkrete regelhafte Festlegungen z. B. von Schwellenwerten für p-Werte begrüßt werden, aber darüber hinaus gefordert wird, auch jede Ausnahme von der Regel klar zu definieren und festzulegen. Das Institut stellt mit dem Methodenpapier ein Regelwerk vor, das es erlaubt, in transparenter Weise methodische Bewertungen vorzunehmen. Hierzu sind zum Teil arbiträre Festlegungen von z. B. Schwellenwerten notwendig, die sich nicht exakt begründen lassen, wie z. B. die Festlegung, dass zum Nachweis von Effekten in der Regel ein Signifikanzniveau von 0,05 gewählt wird. Andererseits ist es aber auch nicht möglich und auch nicht Sinn eines Regelwerks, jede Ausnahme von der Regel in klar operationalisierter Weise zu definieren und darzustellen.

2.2 Würdigung der Stellungnahmen zu Abschnitt 2.1.1 „Bericht" und Abschnitt 2.2.3 „Review der Produkte des Instituts"

In mehreren Stellungnahmen wird die Umwandlung des bisher verpflichtend durchgeführten externen Reviews im Rahmen der Berichtserstellung in einen optionalen Schritt kritisiert.

Für die Qualitätssicherung von Produkten als IQWiG-interne Aufgabe ist im Institut ein umfangreiches internes Qualitätssicherungssystem etabliert, innerhalb dessen Produkte vor Veröffentlichung einem mehrschrittigen Reviewverfahren unterzogen werden. Erfahrungen der letzten Jahre haben gezeigt, dass die im Zusammenhang der Produkterstellung (teils optional) durchgeführten externen Reviews auch im Rahmen der Erstellung von Berichten nicht regelhaft notwendig sind. Eine Notwendigkeit für die bisher regelhafte Beauftragung eines Reviews des Vorberichts wird umso weniger gesehen, als für Externe grundsätzlich die

Möglichkeit zur Stellungnahme (einschließlich der Veröffentlichung und Würdigung derselben) besteht und diese rege genutzt wird.

Das Institut kommt darüber hinaus unverändert seinem gesetzlichen Auftrag zur Einbeziehung externer Sachverständiger durch die regelhafte Beteiligung externer Sachverständiger nach.

2.3 Würdigung der Stellungnahmen zu Abschnitt 3.1.4 „Endpunktbezogene Bewertung“

In einer Stellungnahme wird empfohlen, den Begriff „Beleglage“ zu modifizieren, da es zu Irritationen mit der Aussagesicherheitskategorie „Beleg“ geben kann.

Es wurde entschieden, den Begriff „Beleglage“ nicht zu modifizieren, da er vom Institut seit Langem benutzt wird. Es wurden stattdessen Erläuterungen ergänzt, sodass mögliche Irritationen vermieden werden.

In mehreren Stellungnahmen gab es Vorschläge zur Struktur und Beschriftung von Tabelle 2 zu den abgeleiteten Aussagesicherheiten für verschiedene Evidenzsituationen.

Die Vorschläge zur Tabelle 2 wurden vom Institut aufgegriffen und die Struktur und Beschriftung von Tabelle 2 modifiziert.

In einer Stellungnahme wird vorgeschlagen, die beiden beschriebenen Arten von Ergebnissicherheiten (qualitative und quantitative Ergebnissicherheit) zunächst durch gängige epidemiologische Begriffe zu definieren.

Dieser Vorschlag wurde umgesetzt.

In einer Stellungnahme wird beschrieben, dass es nicht ganz klar ist, dass sich die qualitative Ergebnissicherheit auf die Studien- und die Endpunktebene bezieht.

Es wurde durch entsprechende Ergänzungen verdeutlicht, dass sich die qualitative Ergebnissicherheit auf die Studien- und die Endpunktebene bezieht.

In mehreren Stellungnahmen wird das Institut aufgefordert, die Kriterien zu ergänzen, wann aus nur 1 Studie Belege abgeleitet werden können.

Dieser Vorschlag wurde im Institut sowie in der wissenschaftlichen Erörterung am 30.09.2013 intensiv diskutiert. Als Ergebnis dieser Diskussionen wurden die Kriterien, wann aus nur 1 Studie in Ausnahmefällen Belege abgeleitet werden können, im Methodenpapier ergänzt.

In einer Stellungnahme wird gefordert, dass in der frühen Nutzenbewertung eine Anpassung der Aussagesicherheit erfolgen sollte, da hier regelhaft nur eine geringe Zahl an Studien – häufig sogar nur 1 Studie – vorliegt.

Das Institut hält dies für ein unlogisches Argument. Die Tatsache, dass nur geringe Evidenz vorliegt oder vorliegen kann, erhöht ja nicht die Aussagesicherheit. Es ist die wissenschaftliche Aufgabe des Instituts, die Aussagesicherheit gemäß der vorhandenen Evidenzlage darzustellen. Ob bei unterschiedlicher Aussagesicherheit darauf basierende Entscheidungen aufgrund anderer Argumente gleich ausfallen oder nicht, ist eine andere Frage.

In einer Stellungnahme wird vorgeschlagen, im Methodenpapier zu ergänzen, ob es möglich ist, die Evidenzlage, die sich aus direkten Vergleichsstudien ergibt, durch indirekte Vergleiche aufzuwerten.

Diese Möglichkeit gibt es prinzipiell und es wurden entsprechende Ergänzungen im Methodenpapier vorgenommen.

2.4 Würdigung der Stellungnahmen zu Abschnitt 3.1.5 „Zusammenfassende Bewertung“

In einer Stellungnahme wird angeraten, die Diskussion zu den qualitätsadjustierten Lebensjahren (QALYs) umfassend darzustellen. Der Stellungnehmende äußert die Befürchtung, in der Kürze der Darstellung, der Auswahl der Quellen und dem Verweis auf „ethische und methodische Probleme“ der häufig verwendeten QALYs würde die entsprechende Diskussion verkürzt wiedergegeben.

Das Institut ist an einer umfangreichen wissenschaftlichen Diskussion der QALYs als Maß des Gesamtnutzens zum Vergleich von Interventionen interessiert. Allerdings kann im Rahmen des Methodenpapiers nicht die gesamte wissenschaftliche Diskussion zu einzelnen Themenbereichen erschöpfend wiedergegeben werden, zumal dies schon anderswo ausführlich dargestellt wurde. In diesem Zusammenhang sei daher auf die Darstellung zum Einsatz der QALYs in der Kosten-Nutzen-Bewertung in den weiterhin gültigen „Allgemeinen Methoden zur Bewertung von Verhältnissen zwischen Kosten und Nutzen“ Version 1.0 vom 12.10.2009 verwiesen. Dort heißt es auf S. 19: „Die indikationsspezifische Anwendung von QALYs kann insbesondere dann sinnvoll sein, wenn es sich um neue Arzneimittel handelt, deren lebensverlängernde Wirkung mit hohen Einbußen der Lebensqualität durch Nebenwirkungen erkauft wird.“ Das Institut kann also QALYs in Kosten-Nutzen-Bewertungen (KNBs) nutzen.

Auch andere Stellungnehmende haben betont, dass „*[d]ie intensive Erforschung und Diskussion des QALY-Konzepts […] die ihm zugrundeliegenden Werturteile, Annahmen und Defizite deutlich*“ gemacht hätten. Das heißt, die ethischen und methodischen Probleme sind in der wissenschaftlichen Fachwelt hinlänglich bekannt. Das Institut weist in diesem Zusammenhang darauf hin, dass das QALY-Konzept außer in den von einem Stellungnehmer zitierten Passagen der Stellungnahme des Deutschen Ethikrats zu „Nutzen und Kosten im Gesundheitswesen – Zur Normativen Funktion ihrer Bewertung“ (S. 37 ff.) auch an anderen Stellen der Stellungnahme des Deutschen Ethikrats beleuchtet wird (S. 61–70). So wird auf den Seiten 37 ff. zunächst die Erhebung des QALY einfach nur beschrieben. In der

nachgelagerten Passage (S. 61–70) der Stellungnahme des Deutschen Ethikrats werden dann die methodischen und ethischen Aspekte ganz allgemein und übergreifend diskutiert. Der Deutsche Ethikrat kommt abschließend zu folgender Einschätzung des QALY-Konzepts [2]: „Insgesamt zeigen die vorstehenden Ausführungen, dass jede QALY-Korrektur zugunsten bestimmter Personengruppen auf bestimmten Wertprämissen beruht, die als solche nicht selbstverständlich sind und in ihrer Gewichtung untereinander problematisch sein können. Die Probleme des QALY-Konzepts sind wegen der fraglichen Konzeption und wegen der unsicheren Berechnungsgrundlage evident."

Inhaltlich sieht das Institut keinen Änderungsbedarf. Die Stelle wurde aber im Abgleich mit den „Allgemeinen Methoden zur Bewertung von Verhältnissen zwischen Nutzen und Kosten" angepasst. Weiterhin wurden Angaben zu Literatur, in der sich sowohl Argumente für als auch wider das QALY-Konzept finden, ergänzt. Darunter befinden sich auch die von einem Stellungnehmenden zitierten Quellen zum QALY als ein Instrument, mit dem Nutzen und Schaden aggregiert ausgedrückt werden können. Allerdings muss eine der Quellen, die ein Stellungnehmer anführt (Puhan et al. [7]), kritisch diskutiert werden. Die genannte Quelle von Puhan et al. [7] geht nur sehr wenig auf die Probleme ein, wenn QALYs zur Abwägung von Nutzen und Schaden herangezogen werden. Die Frequenz der Erhebung und die Änderungssensitivität der verschiedenen Erhebungsmethoden z. B. werden nicht ausreichend kritisch gewürdigt.

In einigen Stellungnahmen wird angeraten, die Methoden Analytic Hierarchy Process (AHP) und Conjoint Analyse (CA) weiter zu prüfen und zu untersuchen, ob damit Präferenzen erhoben werden können. Im Detail gehen unterschiedliche Stellungnehmende auf folgende Problembereiche bzw. Fragenkomplexe dieser beiden methodischen Ansätze ein: 1) Repräsentativität bzw. Auswahl der zu Befragenden, 2) Validität und Vergleichbarkeit der Ergebnisse, 3) Stärken und Schwächen vor dem Hintergrund der unterschiedlichen mathematischen Grundlagen und der unterschiedlichen Fundierung in der ökonomischen Theorie.

Das Institut regt an, wie in der mündlichen Erörterung am 30.09.2013 erläutert und auch von einigen Stellungnehmenden so geäußert, den Diskussionsprozess zu verschieben, da nach Publikation eines Arbeitspapiers zu einem Pilotprojekt AHP (Mai 2013) demnächst auch ein Arbeitspapier zu einem Pilotprojekt CA erscheint. Es ist geplant, die Ergebnisse und Erkenntnisse aus diesen Pilotprojekten bei der Neufassung entsprechender Abschnitte des Methodenpapiers einfließen zu lassen. Diese Abschnitte würden dann in einem Stellungnahmeverfahren zur Diskussion gestellt.

2.5 Würdigung der Stellungnahmen zu Abschnitt 3.3.3 „Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V“ sowie zum Anhang „Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens“

In mehreren Stellungnahmen wird gefordert, dass die in der Arzneimittel-Nutzenbewertungsverordnung zur Quantifizierung des Ausmaßes des Zusatznutzens aufgeführten Begriffe „Heilung“ und „spürbare Linderung der Erkrankung“ als eigene Entitäten bestehen bleiben müssten.

Die im Anhang zum Entwurf enthaltene Rationale widmet sich auf den ersten Seiten ausführlich den gesetzlichen Vorgaben und der daraus abgeleiteten Methodik zur Feststellung des Ausmaßes des Zusatznutzens. Dort ist ersichtlich, dass die Vorgaben zum Teil eindeutig, zum Teil aber auch weniger eindeutig und sogar lückenhaft sind. Die Vorgaben in der Arzneimittel-Nutzenbewertungsverordnung sind demnach eher exemplarisch als erschöpfend anzusehen. Daher galt es, anhand der Vorgaben eine Operationalisierung zu entwickeln. So wird z. B. in der Rationale erläutert, dass der Begriff „Heilung“ einer Operationalisierung bedarf und dazu Kriterien anzulegen seien, die sich auch in den Endpunkten Mortalität, Morbidität und Lebensqualität abbilden lassen. Es wird in diesen Stellungnahmen leider kein Vorschlag unterbreitet, wie die Begriffe „Heilung“ und „spürbare Linderung der Erkrankung“ anders als beschrieben in die Methodik eingebettet werden könnten.

In einer Stellungnahme wird mit Verweis auf Abschnitt 3.3.3 gebeten, anstelle einer 4-fachen Abstufung – entsprechend der Verfahrensordnung des G-BA – die „6-fache Abstufung zur Belegbarkeit des (Zusatz-)Nutzens zu übernehmen“.

Hier liegt offensichtlich ein Missverständnis vor. Die 6 Kategorien, die die Arzneimittel-Nutzenbewertungsverordnung vorgibt, beziehen sich auf das Ausmaß des Zusatznutzens in der Gesamtschau. Diese sind auch als solche im Abschnitt 3.3.3 (3. Absatz, S. 15) und im Anhang (2. Absatz, S. 26) genannt. Die in Schritt 2 genannten 4 Kategorien (erheblich, beträchtlich, gering, nicht quantifizierbar) beziehen sich auf die Zwischenschritte der Ausmaßfeststellung auf Endpunktebene. In der Gesamtschau im 3. Schritt finden die 6 in der Arzneimittel-Nutzenbewertungsverordnung genannten Kategorien Anwendung. Zur Vorbeugung dieses Missverständnisses wurden die 6 Kategorien in Schritt 3 des Abschnittes 3.3.3 noch einmal ergänzt.

In einer Stellungnahme wird die Einteilung der Zielgrößen in lediglich 3 Kategorien (Gesamtmortalität, Schwerwiegendes, nicht Schwerwiegendes) aufgrund der „Vielzahl unterschiedlicher Erkrankungen“ kritisiert.

Die Rationale im Anhang erläutert ausführlich die Bildung der 3 Kategorien, die vorrangig aus den gesetzlichen Vorgaben resultieren und zusätzlich aus der Notwendigkeit einer expliziten und abstrakten Operationalisierung. Neben dem unspezifischen Aufruf des Stellungnehmenden, diesbezüglich die Diskussion mit den Fachgesellschaften zu suchen,

enthält die Stellungnahme keinen konkreten Ansatzpunkt oder Vorschlag, wie denn stattdessen vorgegangen werden könnte.

In mehreren Stellungnahmen wird die Methodik zur Feststellung des Ausmaßes des Zusatznutzens mit dem Argument kritisiert, dass der G-BA diese Methodik nicht anerkennt bzw. berücksichtigt.

Es ist richtig, dass sich der G-BA die vom IQWiG verwendete Methodik nicht zu eigen gemacht und z. B. in seine Verfahrensordnung übernommen hat. Faktisch ist der G-BA jedoch in der Mehrzahl der Fälle zur gleichen Ausmaßfeststellung wie das IQWiG gelangt.

In einer Stellungnahme wird die Darstellung im letzten Satz auf Seite 26 („Weiterhin erscheint die Zielgröße (gesundheitsbezogene) Lebensqualität, die in § 2 Abs. 3 der AM-NutzenV explizit als Nutzenkriterium formuliert wird, überhaupt nicht in der Kriterienliste für das Ausmaß des Zusatznutzens.“) als „fehlerhaft“ bezeichnet, da „in § 5 Abs. 7 Nr. 1 bis 3 ... die „Lebensqualität“ durch Zitierung des § 2 Abs. 3 der AM-NutzenV indirekt eingefügt“ werde.

Der in der Stellungnahme kritisierte Satz besagt richtigerweise, dass in der Arzneimittel-Nutzenbewertungsverordnung keine Kriterien *für das Ausmaß des Zusatznutzens* genannt werden, die sich auf die Lebensqualität beziehen. Die Richtigkeit wird nicht dadurch hinfällig, dass vom Stellungnehmenden ein anderer Absatz zitiert wird, der auch keine diesbezüglichen Kriterien enthält. Der vom Stellungnehmenden dargestellte Sachverhalt verdeutlicht aber, dass die in der AM-NutzenV genannten Kriterien zur Feststellung des Ausmaßes des Zusatznutzens nicht lückenlos sind. Dies bekräftigt das vom Institut gewählte Vorgehen, diese beispielhafte Aufzählung zu ergänzen.

Des Weiteren wird in der Stellungnahme dazu aufgefordert, im 2. Absatz auf Seite 28 „der Darstellung der Umstrukturierung die weiteren in der Verordnung genannten Zielgrößen, Überlebenszeit und Krankheitsdauer beizufügen, um einer vollständigen exakten Aufzählung zu entsprechen“.

Die Zielgröße Überlebenszeit ist im 1. Punkt „Gesamtmortalität“ des entsprechenden Absatzes enthalten, was auch aus der Angabe zur „Verlängerung der Überlebensdauer“ in der Spalte „Gesamtmortalität“ der Tabelle NT3 ersichtlich ist. Die Zielgröße „Krankheitsdauer“ wird unter den Kriterien *für das Ausmaß des Zusatznutzens*, um die es in diesem Absatz geht, nicht genannt und ist daher auch nicht zu ergänzen.

In 2 Stellungnahmen wird gefordert, das Vorgehen bei der Ableitung der Gesamtaussage zum Ausmaß des Zusatznutzens genauer zu beschreiben.

Eine Operationalisierung für diesen Schritt der Gesamtschau hält das Institut zurzeit nicht für möglich. Für die Gesamtabwägung existiert keine allgemein akzeptierte Methodik. Auch in den Stellungnahmen werden keine konkreten Vorschläge diesbezüglich unterbreitet. Auf der

anderen Seite ist es Aufgabe des Instituts im Rahmen der frühen Nutzenbewertung, eine Gesamtaussage zum Ausmaß des Zusatznutzens vorzuschlagen, wozu dann im Rahmen des G-BA-Verfahrens Stellung genommen werden kann. In den Dossierbewertungen des Instituts wird die Herleitung der Gesamtaussage zum Zusatznutzen transparent dargestellt: Ableitung der Aussagen auf Endpunktebene, Gegenüberstellung von positiven und negativen Effekten sowie Beschreibung der Abwägung dieser positiven und negativen Effekte. Diese Transparenz unterstützt nicht nur die Entscheidungsfindung des G-BA, sondern ermöglicht auch Stellungnehmenden, sich zu dieser Abwägung gezielt zu äußern und Abweichungen zu begründen. Die abschließende Entscheidung trifft der G-BA. Damit wird den gesetzlichen Anforderungen inhaltlich und prozedural vollumfänglich Rechnung getragen.

Eine Stellungnahme beinhaltet die Forderung nach einer Einbettung in einen wissenschaftlichen Diskussionsprozess und führt aus, dass „nicht nachvollziehbar dargestellt [sei], auf welcher Basis das IQWiG zu einer horizontalen und vertikalen Abstufung hinsichtlich der in der Matrix der Tabelle NT6 festgelegten relativen Risiken und der daraus für die Erreichung der jeweiligen Zielgröße zu unterschreitenden Schwellenwerte gekommen ist“.

Die vorgebrachte Kritik ist unbegründet. Zum einen enthält der Methodenentwurf einen Anhang, in dem auf etwa 12 Seiten ausführlich die vom Stellungnehmenden geforderte „Basis“ zur „horizontalen und vertikalen Abstufung“ für die Feststellung des Ausmaßes des Zusatznutzens beschrieben ist. Vor diesem Hintergrund erscheint die pauschale Kritik ohne Benennung konkreter Punkte nicht hilfreich. Zum anderen stellt die Veröffentlichung des Methodenentwurfs mit dem expliziten Aufruf zur Abgabe von Stellungnahmen, die schriftliche Würdigung dieser Stellungnahmen sowie mündliche Erörterung gerade den vom Stellungnehmenden geforderten wissenschaftlichen Diskussionsprozess dar.

In mehreren Stellungnahmen wird eine „Priorisierung des Endpunkts Mortalität“ über andere Endpunkte wie z. B. Lebensqualität oder schwerwiegende Symptome kritisiert, insbesondere bei Erkrankungen, die nicht lebensbedrohend sind.

Es ist festzuhalten, dass im Entwurf und seinem Anhang an keiner Stelle eine Priorisierung der Zielgrößen Mortalität genannt ist. Vielmehr ist beschrieben, dass die Feststellung des Ausmaßes des Zusatznutzens zunächst für jeden (patientenrelevanten) Endpunkt separat bestimmt wird (siehe Schritt 2 auf S. 16 im Entwurf). Im 3. Schritt der Gesamtschau über die Endpunkte hinweg wird explizit (siehe S. 19, letzter Absatz im Entwurf) auf eine Hierarchisierung der Endpunkte verzichtet.

Mehrere Stellungnehmende kritisieren die falsche bzw. ungenügende Berücksichtigung des Schweregrads der Erkrankung, wie sie in der Arzneimittel-Nutzenbewertungsverordnung gefordert ist.

Leider verzichten die Stellungnehmenden auf Hinweise, an welcher konkreten Stelle im Entwurf der Sachverhalt aus ihrer Sicht falsch bzw. ungenügend umgesetzt wurde. Darüber hinaus muss auch an dieser Stelle festgestellt werden, dass vonseiten der Stellungnehmenden keine konkreten und konstruktiven Verbesserungsvorschläge unterbreitet wurden.

In zwei Stellungnahmen wird gefordert, dass insbesondere für Symptome die Unterscheidung zwischen „schwerwiegend / schwer“ und „nicht schwerwiegend / nicht schwer“ zu operationalisieren sei.

Die Frage, wann ein bestimmter Endpunkt als schwerwiegend oder nicht schwerwiegend einzustufen ist, lässt sich nicht für alle Entitäten global beantworten und muss daher im konkreten Fall entschieden werden. Teilweise wird in den relevanten Studien eine Unterscheidung vorgenommen (z. B. schwerwiegende unerwünschte Ereignisse vs. nicht schwerwiegende unerwünschte Ereignisse, schwere oder nicht schwere Symptome etc.). Teil der Bewertung ist dann zu prüfen, ob diese Einstufung sachgerecht ist. Teilweise liegt eine solche Unterscheidung in den Studien nicht vor. Für die Einstufung kann es notwendig sein, dass hierzu weitere krankheitsspezifische Informationsquellen herangezogen werden. Neben Fachliteratur sind dies z. B. die Patientenperspektive oder die Einschätzung von Fachexperten. Im Methodenpapier des IQWiG ist dargestellt, dass beide Perspektiven regelhaft bei der Dossierbewertung berücksichtigt werden. Darüber hinaus bieten die Dossiervorlagen ausreichend Möglichkeiten für den jeweiligen pharmazeutischen Unternehmer, sich hierzu zu äußern und seine Einschätzung zu begründen. Auch dies kann bei der Einschätzung des Schweregrads eines Symptoms oder einer Nebenwirkung hilfreich sein.

In zwei Stellungnahmen wird gefordert, die auf den Endpunkt Lebensqualität bezogenen „Standards für die Art, Auswahl und Auswertung der verwendeten Instrumente“ zu konkretisieren.

Der Umgang mit Endpunkten zur Lebensqualität ist bereits in der bisherigen Version 4.0 der Allgemeinen Methoden, konkret in den Abschnitten 3.1.1 (S. 31) und 3.2.4 (S. 43), beschrieben. U. a. wird dort auf die relevante Richtlinie der EMA [4] verwiesen, die sich auch zu den (allgemein bekannten) Standards für die Art, Auswahl und Auswertung der verwendeten Instrumente äußert.

In mehreren Stellungnahmen wird die Operationalisierung zur Feststellung des Ausmaßes des Zusatznutzens als zu starr kritisiert. Es wird insbesondere die Notwendigkeit gesehen, anstelle globaler Schwellenwerte eine Differenzierung vorzunehmen, die indikationsspezifisch ist und die Krankheitsschwere berücksichtige. Des Weiteren wird die Verankerung von 0,5 als relatives Risiko für einen Effekt erheblichen Ausmaßes kritisiert. Dieser Anker beruhe nur auf einer Studie und gelte – wenn überhaupt – auch nur für onkologische Indikationen. In diesem Kontext wird auch eine Begründung für die feste Rasterung der für die Ableitung der Schwellenwerte festzulegenden tatsächlichen Effekte gefordert.

Die im Entwurf dargestellte Operationalisierung ist in der Tat abstrakt und unabhängig von der Indikation. Um die Gründe für dieses Vorgehen darzulegen, wurde dem Entwurf eine Rationale angehängt. Sie beschreibt, dass das IQWiG eine Feststellung zum Ausmaß treffen muss, dass dies mit einem möglichst nachvollziehbaren Verfahren erfolgen sollte, dass es hierfür kaum konkrete empirische Grundlagen gibt und dass das Verfahren möglichst wenige Werteentscheidungen (etwa zur unterschiedlichen „Wertigkeit“ von Indikationen) trifft, die dem G-BA aufgrund seiner Legitimation vorbehalten bleiben sollten. Dieser Rahmen führt fast zwangsläufig zu dem vom IQWiG verfolgten Vorgehen. Konkrete Verbesserungs- oder Alternativvorschläge sind dem Institut nicht bekannt.

In einer Stellungnahme wird angeregt, in den Dossierbewertungen neben dem auf relativen Maßen beruhenden Ausmaß des Zusatznutzens auch die „absoluten Ausmaße“ darzustellen. Beispielsweise könne eine 50%ige Reduktion der Mortalität eine Reduktion von 10 % auf 5 % oder ebenso von 1 % auf 0,5 % bedeuten. In diesem Kontext wird in mehreren Stellungnahmen die alleinige Verwendung des relativen Risikos als ungeeignet angesehen.

Im Anhang des Entwurfs (Seite 31 und 32) ist die Wahl für relative Maße zur Feststellung des Ausmaßes des Zusatznutzens begründet. Eine zusätzliche Kategorisierung des Ausmaßes des Zusatznutzens anhand von absoluten Maßen birgt die Gefahr von Inkonsistenzen. Ungeachtet dessen ist die Information, welche Größenordnung die Ereignisraten haben, wichtig bzw. sogar notwendig zur Interpretation der Ausmaße, auch im Hinblick auf die sich anschließende Gesamtabwägung. Daher werden in den Dossierbewertungen zusammen mit den relativen Risiken und daraus abgeleiteten Ausmaßen auch immer die beobachteten Ereignisraten (bzw. medianen Ereigniszeiten bei Überlebenszeitanalysen) pro Gruppe dargestellt. Diese Angaben finden sich jeweils in den Tabellen „Ausmaß des Zusatznutzens auf Endpunktebene“ in Form von „Ereignisanteil: x % vs. y %“. So wird das Ausmaß zwar allein anhand der relativen Maße kategorisiert, die absoluten Effekte werden aber zusätzlich mit abgebildet.

In einer Stellungnahme wird der deutlich erhöhte Umrechnungsaufwand kritisiert, da ggf. Kehrwerte gebildet werden müssten, um Effekte immer als Werte < 1 darzustellen.

Dieses Argument ist nicht nachvollziehbar. Bei relativen Maßen können Effekte unterhalb oder oberhalb von 1 liegen. Zur Vereinfachung der Darstellung der Methodik wurde ohne Beschränkung der Allgemeinheit der Fall von Effekten unterhalb von 1 betrachtet und entsprechende Schwellenwerte präsentiert. Anderenfalls hätten zusätzlich Schwellenwerte für Effekte oberhalb von 1 dargestellt werden müssen (durch Kehrwertbildung). Es ist in beiden Fällen eine (einfache) Kehrwertbildung vorzunehmen, entweder für die Effekte oder für die Schwellenwerte. Ein deutlich erhöhter Umrechnungsaufwand liegt nicht vor.

In einer Stellungnahme wird die Festsetzung, dass vom Effektmaß relatives Risiko ausgehend Zähler und Nenner immer so gewählt werden, dass sich der Effekt (sofern vorhanden) als Wert < 1 realisiert, als nicht ausreichend beschrieben, da die „dem Quotienten innewohnende Asymmetrie“ nicht ausreichend berücksichtigt sei.

Diese Festsetzung dient allein dazu, die Operationalisierung zur Feststellung des Ausmaßes des Zusatznutzens auf Effekte unterhalb von 1 beschränken zu können. Eine Ausweitung auf Effekte oberhalb von 1 würde die Darstellung unnötig verkomplizieren und ist durch einfache Kehrwertbildung überflüssig. Die Asymmetrieeigenschaft des relativen Risikos ist in der Tat durch Kehrwertbildung nicht zu beeinflussen. Diese Eigenschaft wurde daher auch in einem gesonderten Absatz (1. Absatz auf S. 18) im Entwurf adressiert.

In zwei Stellungnahmen wird dazu aufgefordert, eine Begründung für die Übertragung der Schwellenwerte für das relative Risiko auf Hazard Ratios anzugeben.

Im Anhang zum Entwurf der Allgemeinen Methoden Version 4.1 ist ausführlich beschrieben, dass zur Ableitung der Schwellenwerte ausgehend von einem inhaltlich begründeten Anker ein abstrakt-pragmatisches Vorgehen gewählt wurde. Dabei mussten zwangsläufig weitere Annahmen getroffen werden, u. a. dass relative Effektmaße für eine allgemeine Beschreibung von Effektstärken bei binären Daten besser geeignet sind als absolute Effektmaße. Ein wesentlicher Vorteil des relativen Risikos gegenüber der absoluten Risikoreduktion ist, dass keine Festlegungen erfolgen müssen, welche Basisrisiken zugrunde zu legen sind bzw. zu welchen Zeitpunkten Basisrisiken geschätzt werden müssen. Wollte man nun für das Hazard Ratio, dessen inhaltliche Bedeutung dem (epidemiologischen) relativen Risiko (aus einer Vierfeldertafel) sehr ähnlich ist, andere Schwellenwerte festlegen, so müssten genau solche Festlegungen über Basisrisiken erfolgen. Denn grundsätzlich lässt sich zwar – wiederum unter bestimmten Annahmen – ein relatives Risiko in ein Hazard Ratio umrechnen und umgekehrt, aber eben nur unter Zugrundelegung eines Basisrisikos (in der Kontrollgruppe) oder eines festen Zeitpunktes zur Schätzung des Basisrisikos. Dies sollte vermieden werden.

Im Übrigen ließe sich in gleicher Weise allgemein hinterfragen, inwieweit Festlegungen des Signifikanzniveaus (von üblicherweise [zweiseitig] 5 %) von einem Effektmaß auf das andere übertragen werden können. Hier hat man bisher aus guten Gründen auf effektmaßspezifische Festlegungen verzichtet.

Eine Stellungnahme enthält die Überlegung, anstelle fester Schwellenwerte zur Ausmaßfeststellung doch „Kern- und Graubereiche“ einzuführen. Dabei wird auf ein ähnliches Konzept zur Bewertung der klinischen Relevanz verwiesen, welches „vom IQWiG … ins Gespräch gebracht worden ist“.

In der Tat weisen beide Konzepte eine Ähnlichkeit auf. Mit den Begriffen „Kern- und Graubereiche“ bezieht sich der Stellungnehmende auf das Konzept bei der Relevanzbewertung, dass nicht nur eine Relevanzschwelle existiert, die relevante von nicht relevanten Effekten trennt. Vielmehr existiert daneben zusätzlich eine Irrelevanzschwelle, sodass zwischen den beiden Schwellenwerten von einem Graubereich gesprochen werden könnte, für den Effekte weder sicher relevant noch sicher irrelevant sind. Allerdings müssen hier zwei Ebenen getrennt werden, sodass der Begriff Graubereich irreführend sein kann. Die Relevanzschwelle bezieht sich auf einen Punktschätzer. Die darauf basierende Operationa-

lisierung der Relevanzbewertung mündet jedoch in einer Irrelevanzschwelle. Diese ist ein Schwellenwert, der als Bezugsgröße für Konfidenzintervallgrenzen dient. Dieses Vorgehen wird in der bisherigen Version 4.0 der Allgemeinen Methoden im Abschnitt 7.3.3 beschrieben.

Die Ähnlichkeit zur Ausmaßfeststellung besteht darin, dass anstelle einer Relevanzschwelle von einem tatsächlichen Effekt (relatives Risiko von 0,5 beim Endpunkt Gesamtmortalität) ausgegangen wird, um einen Schwellenwert für Konfidenzintervallgrenzen abzuleiten.

Darüber hinaus ist anzumerken, dass in den gesetzlichen Vorgaben keine Graubereiche beschrieben sind und sich auch nicht implizit daraus ableiten lassen.

In einer Stellungnahme wird die Richtigkeit der in Abbildung NA1 des Entwurfs dargestellten Ergebnisse bezweifelt. In dieser Stellungnahme werden eigene Berechnungen aufgeführt, die z. T. deutlich von denen des Entwurfs abweichen. Es wird die Vermutung geäußert, dass dies in unterschiedlichen verwendeten Algorithmen zur Berechnung von Konfidenzintervall-grenzen liegt.

Bei den hier kritisierten Ergebnissen geht es um die Frage, welches tatsächliche relative Risiko eigentlich vorliegen muss, damit die im Entwurf auf Seite 35 abgebildeten Vorgaben erfüllt sind. Um eine mögliche Ungenauigkeit durch Verwendung von auf asymptotischen Ansätzen beruhenden Approximationen zu vermeiden, wurden für den Entwurf Monte-Carlo-Simulationen durchgeführt. Die Berechnungen in der Stellungnahme beruhen auf der asymptotischen Formel von Farrington und Manning [5]. Diese unterschiedlichen Ansätze zur Berechnung erklären jedoch nicht die Diskrepanzen zwischen den Ergebnissen. Ein Vergleich beider Ansätze zeigt sogar eine gute Übereinstimmung. Auch die vom Stellungnehmenden vermutete Verwendung unterschiedlicher Algorithmen zur Berechnung von Konfidenz-intervallgrenzen kann die Diskrepanz nicht erklären (in der Simulation wurde die übliche Normalverteilungsapproximation mit der nach der Deltamethode resultierenden Varianz-schätzung für das Log RR verwendet).

Der Unterschied erklärt sich allein dadurch, dass in den in der Stellungnahme beschriebenen Ergebnissen das tatsächliche relative Risiko im ersten Berechnungsschritt fixiert wurde. Nur im zweiten Berechnungsschritt wurde das tatsächliche relative Risiko variiert. Zur Veranschaulichung ist die konkret im Entwurf verfolgte Fragestellung hier noch einmal für den vom Stellungnehmenden gewählten Ansatz über die asymptotische Formel von Farrington und Manning abgebildet:

1) Fest vorgegeben sei das Signifikanzniveau mit 5 % und die Power mit 90 %, sowohl für die übliche Testung der nicht verschobenen als auch für die Testung der verschobenen Hypothese ($H_0: RR \geq RR_0 \quad vs. \quad H_1: RR < RR_0$).

2) Die Funktion $FN(p_0, RR_1, RR_0)$ ergebe die benötigte Fallzahl nach Farrington und Manning für ein Basisrisiko p_0, ein tatsächliches relatives Risiko RR_1 und die Hypothesengrenze RR_0.
3) Gesucht ist bei gegebenem Basisrisiko p_0 und Schwellenwert RR_0 das tatsächliche relative Risiko RR_1, sodass $FN(p_0, RR_1, RR_0) = 2 * FN(p_0, RR_1, 1)$ gilt.

Der Stellungnehmende hat nun im 3. Schritt neben dem Basisrisiko und dem Schwellenwert auch noch ein tatsächliches relatives Risiko (RR_1^*) für die Testung der nicht verschobenen Hypothese fest vorgegeben und somit das tatsächliche relative Risiko RR_1 gesucht, sodass $FN(p_0, RR_1, RR_0) = 2 * FN(p_0, RR_1^*, 1)$ gilt. Das war jedoch nicht die Intention der Berechnung im Entwurf.

In einer Stellungnahme wird die Art der Rasterung der tatsächlichen Effekte zur Schwellenwertableitung in Frage gestellt. Da Punktschätzer für das relative Risiko logarithmisch normalverteilt seien, käme anstelle einer additiven Rasterung von 1/6 doch eher eine multiplikative Rasterung („d. h. mit konstantem Quotienten benachbarter Effekte“) in Frage. Dieses Vorgehen „hätte insbesondere die extreme Anforderung für schwerwiegende Symptome an einen erheblichen Zusatznutzen abgemildert (gewünschtes RR = 0,31 statt 0,17)“.

Die Verwendung einer multiplikativen Rasterung erscheint aufgrund der Quotientenbildung beim relativen Risiko und der Log-Normalverteilung des Schätzers verständlich. Eine solche Rasterung hätte auch ebenso Verwendung finden können. Die daraus resultierenden Schwellenwerte wären in ganz ähnlicher Größenordnung ausgefallen. Die in der Stellungnahme angegebene multiplikative Rasterung von etwa 0,79 ($\sqrt[5]{0{,}31}$) ergibt gerundet folgende tatsächliche Effekte: 0,31 | 0,39 | 0,49 | 0,62 | 0,79 | 1. Vergleicht man diese mit den Werten, die in der Rationale des Entwurfs (Tabelle NT6, Seite 36) als Spanne der tatsächlichen Effekte für die veranschlagten Schwellenwerte abgebildet sind, ist Folgendes erkennbar: Die großen tatsächlichen Effekte der multiplikativen Rasterung (0,31 und 0,39) sind in der jeweiligen Spanne (0,24 bis 0,38 und 0,34 bis 0,48) enthalten und würden so zu sehr ähnlichen Schwellenwerten führen. Die übrigen tatsächlichen Effekte der multiplikativen Rasterung liegen leicht unterhalb der jeweiligen Spannen, sodass die Verwendung der vorgeschlagenen Rasterung sogar zu kleineren Schwellenwerten (im Sinne höherer Hürden) führte. Für eine Verschärfung der Kriterien sieht das Institut keinen Bedarf.

In einer Stellungnahme wird die fehlende Überprüfbarkeit der „Monte-Carlo-Analyse“ beklagt, da sie „weitgehend undokumentiert und somit auch nicht replizierbar“ sei. Die Stellungnehmenden fordern das Institut auf, Inhalt und Methodik der Monte-Carlo-Simulation genauer zu beschreiben.

Das Institut kommt dieser Aufforderung nach. Die entsprechenden Angaben wurden im Anhang ergänzt.

Mehrere Stellungnehmende kritisieren die zur Ableitung der Schwellenwerte getroffene Annahme, dass zwei Studien für eine Fragestellung vorliegen. Diese Annahme sei in bestimmten Indikationen – z. B. der Onkologie – nicht die Regel. Die Verwendung der auf dieser Annahme beruhenden Schwellenwerte führe somit zu einem Powerverlust, falls nur eine Studie vorliegt.

Es ist richtig, dass nicht für jede Fragestellung zwei Studien vorliegen. Die Replikation wissenschaftlicher Ergebnisse wird jedoch grundsätzlich gefordert. Daher ist die Durchführung von zwei Studien für die Zulassung übliche Praxis [3]. Es ist der entscheidende Punkt in der Ableitung der Schwellenwerte, dass die Schwellenwerte gerade so groß gewählt werden, dass beim üblichen Vorliegen von zwei Studien die Power für eine entsprechende Ausmaßkategorie eben nicht niedriger im Vergleich zur ursprünglich angesetzten Power ist. Es liegt in der Natur der statistischen Verfahren, dass die Power für eine Studie, die zur Testung der klassischen Testung auf einen Unterschied angelegt wird, bei Verschiebung (in den Relevanzbereich) nur kleiner werden kann. Die stärkere Aussage erkauft man sich hier durch einen Powertribut. Es ist somit unmöglich, für die Ausmaßfeststellung dieselbe Sicherheit wie für die Testung auf einen Unterschied zu erhalten, ohne die Fallzahl zu erhöhen.

Darüber hinaus wird auch die Verwendung von Konfidenzintervallgrenzen zur Ausmaßfeststellung kritisiert, die bestimmte Populationen benachteilige. So seien „Studiensettings mit einem großen Vorwissen“ benachteiligt, aufgrund der Vorinformationen so präzise planen können, dass das Ergebnis gerade statistisch signifikant ist (und das Konfidenzintervall gerade den Nulleffekt ausschließt).

Dieses Argument ist nicht nachvollziehbar. Läge für eine Fragestellung wirklich ein solches Vorwissen vor, dass ein Studienergebnis so präzise wie oben beschrieben vorhergesagt werden könnte, wäre eine solche Studie überflüssig. Die Effektstärke wäre hinreichend sicher bekannt. Zusätzlich sei angemerkt, dass die Studienplanung fast immer auf einen primären Endpunkt ausgerichtet ist. Das Institut bewertet jedoch alle in den Studien erhobenen patientenrelevanten Endpunkte, für die eine Planung „auf Kante“ gar nicht erfolgen könnte.

Des Weiteren sei bei Studien, in denen die Patienten der Kontrollgruppe bei „Hinweise[n] auf eine positive Wirkung“ vorzeitig das Prüfpräparat erhalten (Cross-Over-Patienten), das Ergebnis in Richtung Nulleffekt verschoben, wodurch es in diesen Situationen ebenfalls zu einer Benachteiligung kommen kann.

Es muss hier beachtet werden, welcher Endpunkt für das Cross-over-Kriterium herangezogen wird. Klinische Studien sind bisher in der Regel nur darauf ausgerichtet, einen statistisch signifikanten Unterschied für einen Wirksamkeitsendpunkt (z. B. progressionsfreies Überleben) zu zeigen (spezielle Fragestellungen wie z. B. Nichtunterlegenheit seien an dieser Stelle einmal ausgenommen). Falls anhand dieses Endpunkts ein Wechsel der Therapien (Cross-over) festgemacht wird, kann der Effekt für diesen Endpunkt unverschoben geschätzt

werden. Für andere Endpunkte kann der Effekt je nach Menge an Therapiewechslern in der Tat zu einer Verschiebung Richtung Nulleffekt führen. Aufgrund der Arzneimittel-Nutzenbewertungsverordnung ist nun aber das Ausmaß eines Effekts – sprich der Abstand zum Nulleffekt – zu quantifizieren. Somit besteht das Problem durch Therapiewechsler unabhängig von der Art der Methodik zur Ausmaßfeststellung. Selbst die – wie in einzelnen Stellungnahmen vorgeschlagen – Verwendung der Effektschätzer anstelle von Konfidenzintervallgrenzen würde das Problem nicht lösen, da die Effektschätzer ebenso in Richtung Nulleffekt verschoben sind.

Schließlich würden seltene Erkrankungen aufgrund der kleinen Populationen systematisch benachteiligt und daher sei „auch der Punktschätzer mit in die Bewertung des Zusatznutzens" einzubeziehen.

Da die Präzision von Studienergebnissen maßgeblich von der Fallzahl abhängt, sind Ergebnisse für Studien mit seltenen Erkrankungen oft unpräzise. Es ist jedoch unklar, wieso dieser Sachverhalt spezifisch bei der Ausmaßfeststellung zu einer *systematischen Benachteiligung* führen soll. Das Problem einer geringen Präzision bei kleinen Studien stellt sich gleichermaßen bei der üblichen Testung auf statistisch signifikante Unterschiede. Je kleiner eine Studie ist, desto größer muss ein geschätzter Effekt ausfallen, um ein statistisch signifikantes Ergebnis zu erhalten. Der Stellungnehmende unterbreitet keine Vorschläge, wie seine Forderung nach dem Einbezug der Punktschätzer in die Bewertung konkret aussehen könnte. Es besteht das Problem bei der Verwendung der Punktschätzer, dass beide Fehlerarten nicht kontrolliert werden. Zum einen kann die Wahrscheinlichkeit für die fälschliche Attestierung eines Ausmaßes nicht begrenzt werden (bzw. beträgt u. U. 50 %). Zum anderen lässt sich auch die Power nicht kontrollieren. Sie ist u. U. selbst bei Erhöhung der Fallzahl nicht größer als 50 %.

In einer weiteren Stellungnahme wird die „Vermengung des Therapieeffektes mit der Präzision seiner Schätzung" als gravierender Nachteil angesehen.

Dieses Argument kann nicht nachvollzogen werden, da es das zentrale Prinzip der schließenden Statistik darstellt und seit Jahrzehnten in der klinischen Forschung etabliert ist. Die üblichen Teststatistiken geben gerade das Verhältnis der geschätzten Effektstärke zur Präzision (dem Standardfehler) wieder. Das gilt auch für die Statistiken, die bei verschobenen Hypothesen wie z. B. zur Testung auf Nichtunterlegenheit herangezogen werden. Durch diesen Ansatz ist es überhaupt nur möglich, beide Fehlerarten (siehe oben) zu kontrollieren.

In mehreren Stellungnahmen wird das Heranziehen von Responderanalysen zur Feststellung des Ausmaßes des Zusatznutzens bei stetigen Endpunkten kritisiert, da durch die Dichotomisierung stetiger Zielgrößen ein entsprechender Informationsverlust einhergehe.

Es ist richtig, dass durch eine Dichotomisierung ein gewisser Informationsverlust in Kauf genommen werden muss. Auf der anderen Seite erspart man sich eine für jedes Effektmaß

individuelle Spezifizierung von Ausmaßschwellenwerten. Leider ist den Stellungnahmen kein konkreter Verbesserungsvorschlag zur Vermeidung des Informationsverlustes zu entnehmen. In einer Stellungnahme wird die Möglichkeit, relative Effektgrößen wie Cohen's d zu verwenden, genannt. Allerdings, so schreiben die Stellungnehmenden selbst, bliebe aber die Frage, „inwieweit sich diese dann über Indikationen hinweg vergleichen lassen in Bezug auf statistische Signifikanz verknüpft mit klinischer Relevanz und welche Grenzen dann für die Bewertung des Ausmaßes des Zusatznutzens verwendet werden sollen". Diese Frage ist bisher nicht geklärt.

Mehrere Stellungnehmende fordern in diesem Zusammenhang, die Methodik für Zielgrößen, für die keine Responderauswertungen vorliegen, genauer zu beschreiben.

Im Entwurf wird auf Seite 19 beschrieben, dass „im Einzelfall zu eruieren [ist], ob relative Risiken approximiert werden können". An dieser Stelle wird auf die Arbeit von da Costa et al. [1] verwiesen, in der verschiedene Ansätze zur Approximation von Ereignisraten und deren Effektmaßen aufgezeigt werden. Es äußert sich kein Stellungnehmender zu den dort beschriebenen Ansätzen. Es bleibt daher unklar, ob diese Ansätze aus Sicht der Stellungnehmenden geeignet oder eher nicht geeignet sind. Konkrete Vorschläge, wie in dieser Situation sonst vorgegangen werden könnte, finden sich ebenfalls nicht in den Stellungnahmen.

In einer Stellungnahme wird der Standpunkt des Instituts „zur Dichotomisierung am Median der gepoolten Behandlungsgruppen" zur Ausmaßfeststellung erbeten, falls nur stetige Zielgrößen ohne Responderanalysen vorliegen.

Wie in Abschnitt 3.3.3 beschrieben, wird das Institut bei fehlenden Responderanalysen nach Möglichkeit auf Verfahren zur Approximation relativer Risiken zurückgreifen und verweist auf die Arbeit von da Costa et al. [1], die dazu verschiedene Ansätze enthält. Die Wahl des Medians über die Gesamtpopulation für die Verwendung als Cut-off-Wert zur Dichotomisierung ist grundsätzlich möglich. Allerdings ist im konkreten Fall zu eruieren, inwieweit der so gebildete Cut-off-Wert ein aussagekräftiges Responsekriterium darstellt.

In einer Stellungnahme wird angemerkt, dass das im Entwurf beschriebene Vorgehen zur Dichotomisierung bei stetigen Zielgrößen „so nicht in der wissenschaftlichen Literatur beschrieben" sei und „u. a. undefinierte Approximationsverfahren" enthielte.

Diese Anmerkung kann nicht nachvollzogen werden, da im entsprechenden Abschnitt (S. 19) des Entwurfs gerade auf wissenschaftliche Literatur [1] verwiesen wird. In diesem Artikel werden mögliche Approximationsverfahren klar definiert und deren Berechnungsschritte detailliert beschrieben.

In einer Stellungnahme wird die erneute Bewertung „signifikante[r] Unterschiede auf Basis von Responderkriterien" anhand von Schwellenwerten kritisiert. Dies sei eine Abweichung von der bisherigen Vorgehensweise, in der „statistisch signifikante Unterschiede unter Verwendung von akzeptierten Responderkriterien als relevant" angesehen wurden.

Für Skalen mit nicht natürlicher Einheit wird üblicherweise die Relevanz bewertet. Dazu ist in den Allgemeinen Methoden 4.0 [6] im Abschnitt 7.3.3 ein hierarchisches Vorgehen beschrieben, welches auch eine Bewertung anhand von Responderkriterien enthält. Ziel dieser Bewertung ist herauszufinden, ob ein beobachteter Effekt (für Skalen mit nicht natürlicher Einheit) so groß ist, um überhaupt von einem spürbaren Effekt und infolgedessen ggf. von einem Zusatznutzen zu sprechen. Für nicht schwere Ereignisse ist in der AM-NutzenV dargestellt, dass dieser Effekt für das Vorliegen eines geringen Zusatznutzens „mehr als nur geringfügig“ sein soll. In der Methodik des IQWiG ist für nicht schwere Ereignisse (und nur für diese) daher eine Schwelle vorgesehen, die für die Feststellung, ob ein Zusatznutzen (oder höherer Schaden) überhaupt vorliegt, unterschritten werden muss. Dies gilt jedoch für alle nicht schweren Ereignisse, und zwar unabhängig davon, ob sie als Responderanalysen anhand akzeptierter Responderkriterien vorliegen oder nicht (z. B. nicht schwerwiegende unerwünschte Ereignisse).

Das im Entwurf auf Seite 16 abgebildete schrittweise Vorgehen ist also keineswegs eine Abweichung vom bisherigen Vorgehen, sondern logische Folge der Notwendigkeit, das Ausmaß des Zusatznutzens auf Basis der Vorgaben der AM-NutzenV zu bewerten.

2.6 Würdigung der Stellungnahmen zu Abschnitt 7.3.8 Meta-Analysen

In einigen Stellungnahmen wird unterstellt, dass die Anwendung von Prädiktionsintervallen eine Verschärfung der Anforderungen darstellt.

Diese Kritik der Stellungnehmenden beruht offensichtlich auf einem Missverständnis. Prädiktionsintervalle stellen ein wichtiges und objektives Tool dar bei der Bewertung, ob in Studien gefundene Effekte gleichgerichtet sind oder nicht, und wenn ja, ob diese deutlich oder mäßig gleichgerichtet sind. In der Version 4.0 des Methodenpapiers wurde nicht unterschieden zwischen deutlich und mäßig gleichgerichtet und generell waren die Kriterien zur Beurteilung der Gleichgerichtetheit von Effekten nicht so klar operationalisiert wie in der neuen Version 4.1. Die Verwendung von Prädiktionsintervallen ermöglicht es, die Kriterien zur Beurteilung der Gleichgerichtetheit von Effekte so detailliert und transparent wie möglich darzustellen. Dies sollte nicht als Verschärfung der Kriterien missverstanden werden.

In einigen Stellungnahmen gibt es Unklarheit, wieso es sinnvoll ist, in Situationen, in denen aufgrund zu hoher Heterogenität kein gepoolter Effektschätzer dargestellt wird, dennoch ein Prädiktionsintervall zu berechnen.

Das Prädiktionsintervall illustriert das Ausmaß der Heterogenität und ist daher sinnvoll auch in Situationen ohne gepoolten Effektschätzer einsetzbar.

In 2 Stellungnahmen wird eine Reaktion des Instituts auf die Stellungnahme von GMDS und IBS-DR vom Februar 2012 vermisst, in der darauf hingewiesen wird, dass es mittlerweile deutlich bessere Methoden für Meta-Analysen als die bisherigen – auch vom IQWiG angewendeten – Standardmethoden.

Das Institut ist sich dieser Tatsache durchaus bewusst, möchte aber nicht voreilig andere Methoden einsetzen. Es gibt zurzeit eine Arbeitsgruppe der Cochrane Collaboration, an der auch Mitarbeiter des Instituts beteiligt sind, die erarbeiten soll, welche neuen Methoden für Meta-Analysen die bisherigen Standardmethoden ersetzen sollen. Das Institut möchte die Ergebnisse dieser Arbeitsgruppe abwarten, um dann neue verbesserte Methoden für Meta-Analysen gemäß dem aktuellen internationalen Standard einzusetzen.

2.7 Literaturverzeichnis

1. Da Costa BR, Rutjes AWS, Johnston BC, Reichenbach S, Nüesch E, Tonia T et al. Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study. Int J Epidemiol 2012; 41(5): 1445-1459.

2. Deutscher Ethikrat (Ed). Nutzen und Kosten im Gesundheitswesen: zur normativen Funktion ihrer Bewertung; Stellungnahme. Berlin: Deutscher Ethikrat; 2011. URL: http://www.ethikrat.org/dateien/pdf/stellungnahme-nutzen-und-kosten-im-gesundheitswesen.pdf.

3. European Medicines Agency. Points to consider on application with: 1. meta-analyses; 2. one pivotal study [online]. 31.05.2001 [Zugriff: 22.09.2010]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf.

4. 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: 23.04.2013]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003637.pdf.

5. Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat Med 1990; 9(12): 1447-1454.

6. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden: Version 4.0 [online]. 23.09.2011 [Zugriff: 12.11.2013]. URL: https://www.iqwig.de/download/IQWiG_Methoden_Version_4_0.pdf.

7. Puhan MA, Singh S, Weiss CO, Varadhan R, Boyd CM. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. BMC Med Res Methodol 2012; 12: 173.

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	Name	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6
Almirall Hermal GmbH	Krug, Ilona	Ja	Nein	Nein	Nein	Nein	Nein
Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF)	Nothacker, Dr. med. Monika	Ja	Ja	Nein	Ja	Nein	Nein
Arzneimittelkommission der deutschen Ärzteschaft (AkdÄ)	Wille, Dr. med. Hans	Nein	Nein	Ja	Nein	Ja	Nein
Bayer Vital GmbH	Appelrath, Meike	Ja	Nein	Nein	Nein	Nein	Nein
	Meinhardt, Michael	Ja	Nein	Nein	Nein	Nein	Nein
Bristol-Myers Squibb GmbH & Co. KGaA	Höcht, Christina	Ja	Nein	Nein	Nein	Nein	Nein
	Plesnila-Frank, Dipl. Volksw. Carlotta	Ja	Ja	Nein	Nein	Nein	Ja
Bundesarbeitsgemeinschaft Selbsthilfe von Menschen mit Behinderung und chronischer Erkrankung und ihren Angehörigen e. V. – BAG Selbsthilfe	Danner, Dr. Martin	Ja	Nein	Nein	Nein	Ja	Nein
Bundesverband der Arzneimittel-Hersteller e. V. (BAH)	Münster, Prof. Dr. Eva	Ja	Ja	Nein	Nein	Nein	Nein
Bundesverband der Pharmazeutischen Industrie e. V. (BPI)	Gerbsch, Dr. Norbert	Ja	Nein	Nein	Nein	Nein	Nein
	Serrano, Dr. Pablo	Ja	Nein	Nein	Nein	Nein	Nein

Organisation/ Institution	Name	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6
Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e. V. (DGHO)	Oldenburg, Michael	Nein	Nein	Nein	Nein	Nein	Nein
	Wörmann, Prof. Dr. Bernhard	Nein	Nein	Nein	Nein	Nein	Nein
Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS)	Hauschke, Prof. Dr. Dieter	Nein	Ja	Ja	Nein	Nein	Nein
Deutsche Krankenhausgesell-schaft e. V. (DKG)	Bielecki, Dr. Eva	Ja	Nein	Nein	Nein	Nein	Nein
	Neumeyer-Gromen, Dr. Angela	Ja	Nein	Nein	Nein	Nein	Nein
Deutsche Region der Internationalen Biometrischen Gesellschaft (IBS-DR)	Schwenke, Dr. Carsten	Nein	Ja	Ja	Nein	Nein	Nein
Gemeinsamer Bundesausschuss (G-BA)	Teupen, Susanne	Nein	Nein	Nein	Nein	Nein	Nein
GlaxoSmithKline GmbH & Co. KG (GSK)	Henning, PD Dr. Michael	Ja	Nein	Nein	Nein	Ja	Ja
Janssen-Cilag GmbH	Erhardt, Wilma	Ja	Nein	Nein	Nein	Nein	Ja
	Fleischmann, Dr. Jochen	Ja	Nein	Nein	Nein	Nein	Ja
Lundbeck GmbH	Friede, Dr. Michael	Ja	Nein	Nein	Nein	Nein	Nein
	Kessel-Steffen, Markus	Ja	Ja	Nein	Nein	Nein	Nein
MSD Sharp & Dohme GmbH	Krobot, Dr. Dr. med. Karl J.	Ja	Nein	Ja	Nein	Nein	Ja
	Scheuringer, Dr. hum. Biol. Monika	Ja	Nein	Nein	Nein	Nein	Nein
Novartis Pharma GmbH	Neidhardt, Katja	Ja	Nein	Nein	Nein	Nein	Nein
	Sieder, Christian	Ja	Nein	Nein	Nein	Nein	Nein
Pfizer Deutschland GmbH	Kloss, Sebastian	Ja	Nein	Nein	Nein	Nein	Nein
	Leverkus, Friedhelm	Ja	Nein	Nein	Nein	Nein	Ja
Sanofi-Aventis Deutschland GmbH	Rosenfeld, Dr. med. Stephanie	Ja	Nein	Nein	Nein	Nein	Ja
	Schinzel, Stefan	Ja	Nein	Nein	Nein	Nein	Nein
Verband Forschender Arzneimittelhersteller e. V. (vfa)	Dintsios, Dr. Ch.-Markus	Ja	Nein	Nein	Nein	Nein	Nein
	Rasch, Dr. Andrej	Ja	Nein	Nein	Nein	Nein	Nein

Organisation/ Institution	Name	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6
Vinzenzkrankenhaus Hannover gGMbH	Schüttert, Dr. med. Jan Bernd	Nein	Nein	Ja	Nein	Nein	Nein
Wissenschaftlicher Beirat	Köbberling, Prof. Dr. Johannes	Nein	Nein	Nein	Nein	Nein	Ja

3.2 Potenzielle Interessenkonflikte von stellungnehmenden Privatpersonen

Name	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6
Wink, Prof. Dr. med. Konrad	Nein	Nein	Nein	Nein	Nein	Nein

Im „Formblatt zur Offenlegung potenzieller Interessenkonflikte“ wurden folgende 6 Fragen gestellt (Version 12/2011):

Frage 1: Sind oder waren Sie innerhalb des laufenden Jahres und der 3 Kalenderjahre davor angestellt bei einem Unternehmen, einer Institution oder einem Interessenverband im Gesundheitswesen, insbesondere bei einem pharmazeutischen Unternehmen, einem Hersteller von Medizinprodukten oder einem industriellen Interessenverband?

Frage 2: Beraten Sie oder haben Sie innerhalb des laufenden Jahres und der 3 Kalenderjahre davor ein Unternehmen, eine Institution oder einen Interessenverband im Gesundheitswesen, insbesondere ein pharmazeutisches Unternehmen, einen Hersteller von Medizinprodukten oder einen industriellen Interessenverband, direkt oder indirekt beraten?

Frage 3: Haben Sie innerhalb des laufenden Jahres und der 3 Kalenderjahre davor direkt oder indirekt von einem Unternehmen, einer Institution oder einem Interessenverband im Gesundheitswesen, insbesondere einem pharmazeutischen Unternehmen, einem Hersteller von Medizinprodukten oder einem industriellen Interessenverband, Honorare erhalten für Vorträge, Stellungnahmen oder Artikel?

Frage 4: Haben Sie und / oder hat die Einrichtung[1], die Sie vertreten, abseits einer Anstellung oder Beratungstätigkeit innerhalb des laufenden Jahres und der 3 Kalenderjahre davor von einem Unternehmen, einer Institution oder einem Interessenverband im Gesundheitswesen, insbesondere einem pharmazeutischen Unternehmen, einem Hersteller von Medizinprodukten oder einem industriellen Interessenverband, finanzielle Unterstützung für Forschungsaktivitäten, andere wissenschaftliche Leistungen oder Patentanmeldungen erhalten?

Frage 5: Haben Sie und / oder hat die Einrichtung[1], bei der Sie angestellt sind bzw. die Sie vertreten, innerhalb des laufenden Jahres und der 3 Kalenderjahre davor sonstige finanzielle oder geldwerte Zuwendungen (z. B. Ausrüstung, Personal, Unterstützung bei der Ausrichtung einer Veranstaltung, Übernahme von Reisekosten oder Teilnahmegebühren ohne wissenschaftliche Gegenleistung) erhalten von einem Unternehmen, einer Institution oder einem Interessenverband im Gesundheitswesen, insbesondere von einem pharmazeutischen Unternehmen, einem Hersteller von Medizinprodukten oder einem industriellen Interessenverband?

Frage 6: Besitzen Sie Aktien, Optionsscheine oder sonstige Geschäftsanteile eines Unternehmens oder einer anderweitigen Institution, insbesondere von einem pharmazeutischen Unternehmen oder einem Hersteller von Medizinprodukten? Besitzen Sie Anteile eines „Branchenfonds“, der auf pharmazeutische Unternehmen oder Hersteller von Medizinprodukten ausgerichtet ist?

[1] Sofern Sie in einer ausgedehnten Institution tätig sind, genügen Angaben zu Ihrer Arbeitseinheit, zum Beispiel Klinikabteilung, Forschungsgruppe etc.

4 Dokumentation der wissenschaftlichen Erörterung – Teilnehmerliste, Tagesordnung und Protokoll

4.1 Teilnehmerliste der wissenschaftlichen Erörterung

Name	Organisation / Institution / Firma / privat
Appelrath, Meike	Bayer Vital GmbH
Danner, Martin	Bundesarbeitsgemeinschaft Selbsthilfe von Menschen mit Behinderung und chronischer Erkrankung und ihren Angehörigen e. V. – BAG Selbsthilfe
Dintsios, Ch.- Markos	Verband Forschender Arzneimittelhersteller e. V. (vfa)
Fleischmann, Jochen	Janssen-Cilag GmbH
Friede, Michael	Lundbeck GmbH
Hauschke, Dieter	Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS)
Hennig, Michael	GlaxoSmithKline GmbH & Co. KG
Höcht, Christine	Bristol-Myers Squibb GmbH & Co. KGaA
Kessel-Steffen, Markus	Lundbeck GmbH
Kloss, Sebastian	Pfizer Deutschland GmbH
Köbberling, Johannes	Wissenschaftlicher Beirat
Krobot, Karl J.	MSD Sharp & Dohme GmbH
Krug, Ilona	Bundesverband der Pharmazeutischen Industrie e. V. (BPI)
Leverkus, Friedhelm	Pfizer Deutschland GmbH
Meinhardt, Michael	Bayer Vital GmbH
Münster, Eva	Bundesverband der Arzneimittel-Hersteller e. V. (BAH)
Neidhardt, Katja	Novartis Pharma GmbH
Neumeyer-Gromen, Angela	Deutsche Krankenhausgesellschaft e. V. (DKG)
Nothacker, Monika	Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF)
Oldenburg, Michael	Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e. V. (DGHO)
Plesnila-Frank, Carlotta	Bristol-Myers Squibb GmbH & Co. KGaA
Rasch, Andrej	Verband Forschender Arzneimittelhersteller e. V. (vfa)
Rosenfeld, Stephanie	Sanofi-Aventis Deutschland GmbH
Scheuringer, Monika	MSD Sharp & Dohme GmbH
Schinzel, Stefan	Sanofi-Aventis Deutschland GmbH
Schüttert, Jan Bernd	Vinzenzkrankenhaus Hannover gGmbH
Schwenke, Carsten	Deutsche Region der Internationalen Biometrischen Gesellschaft (IBS-DR)
Serrano, Pablo	Bundesverband der Pharmazeutischen Industrie e.V. (BPI)
Sieder, Christian	Novartis Pharma GmbH
Teupen, Susanne	Gemeinsamer Bundesausschuss (G-BA)
Wille, Hans	Arzneimittelkommission der deutschen Ärzteschaft (AkdÄ)
Wink, Konrad	privat

Name	**Organisation / Institution / Firma / privat**
Wörmann, Bernhard	Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e. V. (DGHO)
Beckmann, Lars	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Bender, Ralf	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Gerber, Andreas	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Grouven, Ulrich	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Kaiser, Thomas	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Lange, Stefan	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Skipka, Guido	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Thomas, Stefanie	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Vervölgyi, Volker	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Windeler, Jürgen	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)

4.2 Liste der Stellungnahmen, zu denen kein Vertreter an der wissenschaftlichen Erörterung teilgenommen hat

In der folgenden Tabelle werden Stellungnahmen genannt, zu denen kein Stellungnehmender oder Vertreter zur wissenschaftlichen Erörterung erscheinen konnte.

Organisation / Institution / Firma / Privatperson
Deutsche Gesellschaft für Gesundheitsökonomie e. V. (dggö)
GKV-Spitzenverband
Herescon GmbH
Meyer, Gabriele
Röhmel, Joachim

4.3 Tagesordnung der wissenschaftlichen Erörterung

	Begrüßung und Einleitung
TOP 1	**Verwendung von Konfidenzintervallgrenzen zur Ausmaßfestlegung**
TOP 2	**Verwendung von Ereignissen oder Nichtereignissen für das relative Risiko**
TOP 3	**Kriterien für Belege aus einer Studie**
TOP 4	**Anwendung und Interpretation von Prädiktionsintervallen**
TOP 5	**Evidenzaufwertung durch indirekte Vergleiche**
TOP 6	**Methodik für Meta-Analysen mit zufälligen Effekten**
TOP 7	**Verschiedenes**

4.4 Protokoll der wissenschaftlichen Erörterung

Datum: 30.09.2013, 11:00 bis 14:45 Uhr

Ort: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Im Mediapark 8, 50670 Köln

Moderation: Jürgen Windeler

Moderator Jürgen Windeler: Ich begrüße Sie alle hier im IQWiG zu der Erörterung zu den Änderungen im Methodenpapier. Ich gehe davon aus, dass einige von Ihnen Erörterungen im IQWiG bereits mitgemacht haben und das übliche Prozedere kennen. Für die ist das jetzt Wiederholung, jedenfalls die ersten zwei Minuten. Für die anderen muss ich meine Anfangsbemerkungen machen, einige will ich machen, einige muss ich machen.

Ich fange mit denen, die ich machen muss, an. Die Erörterung dient der Klärung von Fragen, die uns aus den schriftlichen Stellungnahmen offengeblieben sind, die wir diskutieren möchten, wo wir vielleicht Unklarheiten haben. Die Erörterung dient ausdrücklich nicht dazu, dass diejenigen, die schriftliche Stellungnahmen eingereicht haben, diese hier wiederholen und vorlesen sollen. Gehen Sie davon aus, dass wir die alle sorgfältig gelesen haben. Gehen Sie davon aus, dass wir die Punkte, die wir heute nicht auf der Tagesordnung haben, auch verstanden haben oder glauben, verstanden zu haben. Es geht jetzt also nicht darum, noch einmal darzulegen, was Sie schon schriftlich dargelegt haben. Natürlich wird unter dem Tagesordnungspunkt „Verschiedenes“ noch Gelegenheit sein, Dinge anzusprechen, die nicht in den expliziten Tagesordnungspunkten stehen. Aber primär wird es um diese sechs ausgewählten Punkte gehen.

Der zweite Punkt ist: Diese Erörterung wird wie alle Erörterungen im IQWiG aufgezeichnet. Sie sind darauf hingewiesen worden. Dadurch, dass Sie jetzt hier sind, haben Sie der Aufzeichnung und der Veröffentlichung dieser Aufzeichnung zugestimmt, wie das für alle Erörterungen praktiziert wird. Ich kann den- oder diejenigen, die sich jetzt überlegen, das doch nicht veröffentlicht haben zu wollen, was hier passiert oder was er oder sie hier sagt, die weitere Teilnahme an der Erörterung nicht gestatten, weil das bedeuten würde, dass wir das Protokoll nicht veröffentlichen können, was wir natürlich nicht wollen. Aber ich sehe jetzt

keinen aufstehen, und ich sehe auch sonst keine skeptischen Gesichter. Als Konsequenz aus der Aufzeichnung ergibt sich, dass Sie bitte bei jeder Äußerung, nicht nur bei der ersten, wie das manchmal üblich ist, Ihren Namen sagen, damit die Aufzeichnung und auch der Protokollant die Äußerungen richtig zuordnen können.

Wir haben Ihnen eine Tagesordnung zugeschickt, die noch einmal deutlich macht, welche Punkte wir mit Ihnen unter den TOPs 1 bis 6 prioritär bearbeiten wollen. Wie gesagt, TOP 7 ist wie üblich offen für weitere Punkte. Wir haben als groben Rahmen 16 Uhr festgelegt. Ich strebe 15:30 Uhr als Ende der Erörterung an. Dann haben wir noch ein bisschen Luft, auch was die Planung angeht. Aber ich glaube, eine bisschen stringente Abarbeitung tut der Sache und uns allen gut. Wir haben eine Mittagspause von ungefähr einer halben Stunde vorgesehen. Das bedeutet, dass wir für jeden Tagesordnungspunkt im Durchschnitt etwas mehr als eine halbe Stunde Zeit haben. Ich werde auch versuchen, jedenfalls diesen Durchschnitt einigermaßen im Auge zu behalten, und gucken, dass die einzelnen Punkte, die wir geklärt haben möchten, zu ihrem Recht kommen.

Dieses ist die erste mündliche Erörterung zu dem Methodenpapier, auch zu den Vorversionen des Methodenpapiers. Die hat es bisher nicht gegeben. Der Grund ist sehr naheliegend - das haben Sie auch bei den jetzigen Überarbeitungen gesehen -: Bisher hat das IQWiG immer komplette Revisionen seiner Methodenpapiere vorgelegt. Wenn wir da mündliche Erörterungen gemacht hätten, hätte das immer drei Tage gedauert. Das wollten wir nicht, oder man hat entschieden, dass man das so nicht wollte. Wir haben jetzt das Aktualisierungsverfahren geändert, indem wir Module des Methodenpapiers ändern und anpassen und zur Stellungnahme und zur Erörterung stellen. Das bedeutet auch, dass wir uns sowohl in den TOPs, die Sie sehen, als auch unter „Verschiedenes“ auf die Teile des Methodenpapiers beschränken werden, die jetzt geändert worden sind und die zur Stellungnahme veröffentlicht worden sind. Es geht also nicht um andere Punkte des Methodenpapiers, und es geht auch nicht um das Methodenpapier als solches. Wir werden das Verfahren dieser modularen abschnittsweisen, ein bisschen auch am Bedarf orientierten Änderung weiter so praktizieren. Zu anderen Punkten wird es zu einem späteren Zeitpunkt Gelegenheiten geben, sich sowohl schriftlich als auch, wenn man möchte, mündlich zu äußern.

Gibt es Unklarheiten oder Fragen zu dem jetzt von mir skizzierten und absehbaren Ablauf? - Das sehe ich nicht.

Ich möchte diejenigen, die am Ende des Tisches sitzen, bitten, sich besonders deutlich zu melden, da Sie ein bisschen aus meinem Blickfeld sind. Aber das wird Ihnen bestimmt gelingen.

Wir kommen zum

Tagesordnungspunkt 1:

Verwendung von Konfidenzintervallgrenzen zur Ausmaßfestlegung

Guido Skipka: Guten Morgen, zusammen. Mein Name ist Guido Skipka. Ich bin Biometriker im IQWiG.

Zum Inhaltlichen: Wie Sie in unserem Entwurf gesehen haben, verwenden wir Konfidenzintervallgrenzen, um das Ausmaß festzustellen. Daran ist in den Stellungnahmen Kritik geübt worden.

Lassen Sie mich kurz skizzieren, wie der Ablauf ist. Ich bin jetzt nicht bei der Gesamtabwägung, sondern auf Endpunktebene. Die Gesamtabwägung ist ein separater Schritt. Wir betrachten erst einmal die Endpunkte separat. Lassen Sie mich davon ausgehen, dass wir das Ergebnis für einen Endpunkt vorliegen haben.

Wir haben im Entwurf ein zweischrittiges Vorgehen beschrieben, wie wir das Ausmaß feststellen. Im ersten Schritt schauen wir, ob überhaupt ein Effekt vorliegt. Da wenden wir die üblichen statistischen Signifikanztests an, die jeder von uns kennt. Sie wissen, dass das damit kongruent ist, dass man schaut, ob das Konfidenzintervall den Nulleffekt ausschließt. Wenn dies der Fall ist, gehen wir in einen zweiten Schritt und versuchen, das Ausmaß festzustellen. Wir haben da drei Kategorien: gering, mäßig und erheblich. Auch da - so sieht das unser Entwurf vor - verwenden wir die Konfidenzintervallgrenzen. Aus meiner Sicht ist das eine natürliche Erweiterung des ersten Schritts.

An diesem Vorgehen ist Kritik geübt worden. Es gab generelle Kritik, dass dieses Verfahren abhängig von der Varianz bzw. der Fallzahl sei. Das ist nicht das, was wir hier unter TOP 1 diskutieren wollen. Dass die Sachen von der Fallzahl abhängen, ist ein zentraler Punkt der schließenden Statistik. Das werden wir nach unseren heutigen Standards nicht anders machen können. Wir können das gerne später diskutieren, aber das ist im Moment nicht der zentrale Punkt, der uns wichtig ist. Wichtig sind uns zwei andere Argumente, die genannt wurden, und die möchte ich Ihnen vorstellen.

Das erste kam vom BPI, der sagte, dass bestimmte Populationen durch die Verwendung von Konfidenzintervallen benachteiligt seien, weil Studiensettings mit großem Vorwissen aufgrund der Vorinformation, die man hat, so präzise geplant werden können, dass das Ergebnis gerade statistisch signifikant ausfällt, sprich das Konfidenzintervall gerade den Nulleffekt ausschließt. Mit dem Argument haben wir ein paar Probleme. Wir bewegen uns im Kontext der frühen Nutzenbewertung. Da ist es häufig so, dass wir eben keine umfangreichen Vorinformationen haben. Dazu kommt, dass Studien in der Regel auf einen primären Endpunkt geplant sind. Wer sich mit uns schon beschäftigt hat, weiß, dass das nicht unbedingt immer der zentrale Endpunkt ist, der uns interessiert. Wir betrachten patientenrelevante Endpunkte, auf die die Studien von ihrer Fallzahl und so gar nicht unbedingt angelegt sind. Sie sehen, wir haben ein gewisses Problem mit diesem Argument und möchten Sie noch einmal bitten, im Kontext dieser frühen Nutzenbewertung Ihr Argument vorzustellen.

Moderator Jürgen Windeler: Gibt es jemanden, der dazu etwas erläutern kann?

Pablo Serrano: Können wir das zurückstellen? Ich müsste noch einiges… Ginge das?

Moderator Jürgen Windeler: Dann müssten wir den ganzen TOP 1… Sie sehen mich etwas ratlos, weil es das erste Mal ist in zehn oder 15 Erörterungen, die ich moderiert habe, dass eine solche Frage gestellt wird. Ich frage mich ein bisschen, was passieren wird, wenn wir den Punkt zurückstellen. Ich frage Sie das. Sie sind im Moment nicht antwortfähig, so verstehe ich das.

Ilona Krug: Ich werde gleich dazu eine Stellungnahme abgeben. Wenn wir das zurückgestellt haben, gebe ich eine dezidierte Stellungnahme ab.

Moderator Jürgen Windeler: Ich bleibe dabei, das etwas ungewöhnlich zu finden. Wir stellen die erste Frage von TOP 1 zurück. - Herr Skipka thematisiert den zweiten Punkt.

Guido Skipka: Im Kontext von TOP 1 gibt es noch ein zweites Argument, was ich gerne vorstellen möchte. Das Argument kam von der Firma Bayer Vital und bezieht sich auf seltene Erkrankungen. In dem Argument wird gesagt, dass seltene Erkrankungen aufgrund der kleinen Population systematisch benachteiligt seien und von daher auch Punktschätzer mit in die Bewertung des Ausmaßes einzubeziehen seien.

Ein paar Anmerkungen dazu: Es ist vollkommen klar, dass die Präzision maßgeblich von der Fallzahl abhängt. Das heißt, Studien über seltene Erkrankungen sind naturgemäß unpräziser als Studien über häufige Erkrankungen. Das ist aus unserer Sicht aber kein spezielles Problem der Methodik der Ausmaßfestsetzung. Das ist ein generelles Problem, dass Studien seltener Erkrankungen unpräzise sind. Auch für die klassischen Hypothesentests hat man dieses Problem. Die Forderung nach Punktschätzern ist auf den ersten Blick verlockend, hat aber aus unserer Sicht gravierende Nachteile. Das würde ich gerne mit Ihnen besprechen.

Es wird gesagt, man könnte das Ausmaß auch so festlegen, dass man Schwellenwerte für Punktschätzer anlegt und guckt, ob ein Punktschätzer eine gewisse Schwelle überschreitet. Wenn man etwas näher darüber nachdenkt, wird man sehr schnell feststellen, dass unsere übliche Kontrolle der Fehlerarten - konkret des Alpha- und Beta-Fehlers - aus dem Ruder läuft. Konkret: Der Alpha-Fehler geht auf 50 % herauf. Das kann man sehr leicht zeigen. Paradoxerweise oder unglücklicherweise geht damit noch nicht einmal ein Powergewinn einher. Auch die Power ist unter Umständen auf 50 % begrenzt. Dazu kommt, dass man, wenn man die Fallzahl erhöht, je nach Situation sogar noch eine abnehmende Power hat. Sie sehen, das sind sehr ungünstige Eigenschaften, die wir normalerweise nicht in der Statistik haben. Ich möchte Bayer Vital bitten, diesen Kontext, speziell die Forderung, die Punktschätzer einzubeziehen, noch einmal zu begründen.

Michael Meinhardt: Es war ja nur ein allgemeiner Punkt. Wie Sie schon erwähnten, ist die grundsätzliche Problematik, dass bei seltenen Erkrankungen natürlich weniger Patienten eingeschlossen werden können und dann die Konfidenzintervalle dementsprechend breiter werden. Gerade im Hinblick auf onkologische Indikationen wollten wir auf das Problem allgemein aufmerksam machen. Aber wir teilen die Auffassung, dass es schwierig ist, für

Punktschätzer Schwellenwerte festzulegen, zu definieren. Von daher ist das für uns in dem Sinne okay.

Moderator Jürgen Windeler: Gibt es weitere Bemerkungen, Kommentare, Anmerkungen zu diesem Punkt?

Stefan Schinzel: Es ist doch unbestritten, dass Sie bei dem Vorgehen, das Sie mit der Obergrenze des Konfidenzinterwalls festgelegt haben, eine Vermengung in Kauf nehmen zwischen dem Punktschätzer und der Präzision, mit der er geschätzt wird. Das heißt, Sie können unter Umständen ein Präparat haben, das keinen so großen Zusatznutzen hat, aber eine große Fallzahl kompensiert das hin zu einem gleichen Ausmaß des Zusatznutzens. Ich frage mich, ob damit im Grunde genommen aus der medizinischen Sicht verschiedene Dossiers vergleichbar sind.

Guido Skipka: Herr Schinzel, die Frage ist vielleicht berechtigt. Es ist nun einmal so. Das ist die Frage, die ich habe. Wir machen hier ja nichts anderes als das, was wir seit 30 Jahren machen, indem wir in der klinischen Forschung statistische Hypothesen testen. Genau da passieren diese Dinge. Wenn man sehr präzise ist, kann man natürlich auch sehr kleine Effekte zur Signifikanz bringen. Es gibt ja auch eine jahrzehntelange Diskussion um klinische Relevanz und statistische Signifikanz. Die Probleme sind da. Sie sind meiner Meinung nach kein spezielles Problem der Ausmaßfeststellung. Da ist mir nicht klar, wieso wir jetzt hier an der Stelle dieses Problem gesondert lösen müssen, mit dem wir eigentlich seit 30 Jahren umgehen müssen.

Ch.-Markos Dintsios: Herr Skipka hat natürlich Recht. Wir erfinden nicht die Statistik neu. Nur war es auch so, bevor sich das IQWiG mit der früheren Nutzenbewertung befassen musste, hat es das gleiche statistische Instrumentarium angewendet. Aber die Aussage war eine dichotome Aussage. Effektschätzer und das Konfidenzintervall, in einigen Fällen statistisch signifikant oder nicht signifikant und in einigen Fällen gab es noch die Überlegung hinsichtlich der klinischen Relevanz, wo man dementsprechend eine minimal important clinical difference noch als Kriterium angewendet hat. Jetzt gehen Sie einen Schritt weiter. Sie wenden dasselbe Instrumentarium, dieselben Methoden an und wollen klassifizieren in einem System, das zumindest auf der Positivseite des Zusatznutzens vier Kategorien kennt, so wie es der Gesetzgeber sehen wollte, im Rahmen der Verordnung auch noch festgehalten. Das ist nun einmal ein qualitativer Unterschied. Und der ist nicht Ohne. Ich würde dann die Überlegung anstellen, ob das, was man 30, 40 Jahre gemacht hat, zur Beantwortung der jetzigen Frage führt, ob das, was jetzt angesetzt wird, zielführend ist oder nicht.

Um das aufzugreifen, was Bayer kommentiert und wozu Herr Meinhardt sich positioniert hat: Es stellt sich die Frage, ob man dann für solche Fälle einen anderen Modus finden müsste. Denn letztendlich ist ja nicht Zielsetzung der ganzen Veranstaltung, sich den statistischen Methoden zu unterwerfen, sondern diese weiterzuentwickeln, um die Forschungsfragen so gut es geht zu beantworten.

Guido Skipka: Markos, wir dürfen uns duzen, wir kennen uns. Es ist schon interessant. Wir sitzen heute zusammen, um gegebenenfalls einen neuen Modus zu finden. Die Frage ist natürlich: Wie könnte dieser Modus aussehen?

Stefan Lange: Das ist der eine Punkt.

Der andere Punkt ist: Sie waren jetzt nicht auf das Argument eingegangen. Wenn man sich an Punktschätzer orientieren will, warum will man jetzt eigentlich die Nachteile in beide Richtungen in Kauf nehmen, sowohl das Alpha geht rauf als auch das Beta wird höher, also beide Fehlerarten steigen? Das ist doch nicht besonders sinnvoll. Da wird man doch der besonderen Problematik erst Recht nicht gerecht. Also wäre es schon hilfreich, einen etwas anderen Vorschlag zu machen als nur diesen etwas simplen holzschnittartigen.

Stefan Schinzel: Es wäre ja durchaus denkbar, dass man sich einerseits an der statistischen Signifikanz orientiert und sich andererseits den Punktschätzer separat anschaut zur Feststellung des Ausmaßes.

Dieter Hauschke: Das, was Sie mit dem Punktschätzer vorgeschlagen haben: Seinerzeit hatte ich das mit Meinhard Kieser publiziert, und das ist nach wie vor permanent zitiert worden, oftmals leider so, dass es falsch verstanden wurde. Ich habe gesehen, dass die Methodik dazu verwendet wurde, man nimmt den Punktschätzer und das Konfidenzintervall, die Grenze ist gerade oberhalb von Null. Das war nicht Sinn der Publikation. Die Konsequenz war nachher, dass wir uns mehr oder weniger mit dem IQWiG darauf geeinigt haben, auch unter Einhaltung des Fehlers erster Art, dass man sagt, das gesamte Konfidenzintervall muss eine Schranke überschreiten. Das ist, finde ich, analog zur Statistik. Das ist ein Test. Das ist die Kongruenz zwischen statistischem Test und Konfidenzintervall.

Guido Skipka: In Ergänzung zu dem, was Dieter Hauschke sagt: Wir hatten ja diese Diskussion bei der Relevanzbewertung. Man muss hier natürlich beachten, wenn wir über Schwellenwerte für Punktschätzer und Schwellenwerte für Konfidenzintervallgrenzen reden, dass das nicht dieselben Schwellenwerte sind. Als wir über die Relevanzbewertung geredet und diese Vorschläge diskutiert haben, haben wir uns dem Problem so genähert, dass wir sagen konnten - darauf hatten wir uns auch in großer Runde geeinigt; das war sehr erfreulich - , dass, wenn wir Schwellenwerte für Punktschätzer haben, man da eher von einer Relevanzschwelle redet, dass man aber, wenn wir aber auf Konfidenzintervallgrenzen gehen, eine niedrigere Hürde in Form einer Irrelevanzschwelle braucht. Das ist bei dem Ganzen natürlich zu beachten.

Friedhelm Leverkus: Wenn ich mich richtig erinnere, Guido, ist dieses Schwellenwertmodell abgeleitet worden unter der Voraussetzung, dass man zwei Studien hat und die im Prinzip poolt. Das führt dazu, dass man keinen Powerverlust hat und man kann im Prinzip den Effekt nachweisen. Liegt jedoch nur eine Studie vor, müsste man, um nicht die Null zu erreichen - man plant ja in der Regel auf Null, dass man die Null erreicht -, eine viel größere Fallzahl erreichen, in die Studie einschließen, als geplant worden ist. Habt ihr schon einmal überlegt, ob für diese Situation, wenn nur eine Studie vorhanden ist, dass man wieder die gleichen Bedingungen hat wie bei den zwei Studien?

Guido Skipka: Konkrete Antwort: Wir haben dazu keine Überlegungen. Wir halten sie aber auch nicht für notwendig. Wir hatten die Aufgabe, eine Operationalisierung festzulegen. Wir haben gesagt, wir nehmen Konfidenzintervallgrenzen aus den eben genannten Gründen. Wir brauchen Schwellenwerte. Wie bekommen wir diese Schwellenwerte? Wir wollten jetzt nicht irgendetwas aus dem Bauch heraus festlegen. Wir wollten auch den Realitäten Rechnung

tragen. Üblicherweise haben wir zwei Zulassungsstudien und haben uns daher so genähert, um auch zu gucken. Ich finde nach wie vor, es hat einen ziemlichen Charme, dass man sagen kann, die Power, die man für zwei Zulassungsstudien anlegt, ist genauso groß, wenn man in die Ausmaßfestlegung geht. Das hängt natürlich von der Größe der tatsächlichen Effekte ab.

Wenn wir jetzt in der Situation sind, dass nur eine Studie vorliegt. Ich sage immer salopp: Es gibt nichts umsonst im Leben. Wenn wir neben der einfachen Signifikanzaussage eine stärkere Aussage in Form eines gewissen Ausmaßes machen wollen, die Evidenz ist so, wie sie ist. Und wenn sie für die schwächere Aussage gepowert ist, muss die Power niedriger sein für eine stärkere Aussage. Das können wir natürlich nicht ändern. Ich wüsste auch nicht, wie man dem Rechnung tragen sollte, außer dass man damit leben muss.

Friedhelm Leverkus: Das ist doch nicht so ganz richtig. Die Anzahl der Studien wird ja im Prinzip von den Zulassungsbehörden und solchen Dingen determiniert. Zum einen.

Zum zweiten ist es so: Wenn nur eine Studie vorliegt, findet eine Bewertung ja auch über den Anhaltspunkt oder den Hinweis statt. Man kann keinen Beleg mehr bekommen, sondern letztlich nur einen Hinweis. Das heißt, wenn man im Indikationsgebiet arbeitet, wo letztlich nur eine Studie möglich ist, wird das halt auf zwei Seiten angerechnet, einmal beim Ausmaß und zum anderen bei dem Anhaltspunkt. Da, denke ich, sind schon Überlegungen wert, wie man das wieder richtigstellt.

Ch.-Markos Dintsios: Direkt dazu, ich ergänze nur, denn Herr Leverkus hat es aufgegriffen. Einmal geht es ja über die Konfidenzintervallobergrenzenschwelle. Und dann haben wir noch das Problem bei der einzelnen Studie, dass sie über die Wahrscheinlichkeit, sozusagen in der Trichotomie die Sie haben - wenn ich „Sie“ sage, meine ich das ganze IQWiG; ich kenne ja alle Mitarbeiter des IQWiG, weil ich selber ja lange Zeit beim IQWiG war -, Ausmaß, Hinweis, Beleg. Die Frage, die sich stellt, ist, getriggert wird das zwar über eine Hypothesenannahme, diesen Hypothesenshift, nur da wird etwas als Rahmenbedingung außen vor gelassen. Die Auswahl, eine Studie zu machen und sie über die Zulassungsbehörden entweder in den USA oder in Europa, weil das hauptsächlich unsere geografischen Entitäten sind, wo wir die Zulassung betreiben, seltener in Japan vielleicht, wenn wir sie einreichen, ist nicht unbedingt alleine vom Hersteller zu verantworten. Ich hatte damals in der Stellungnahme des vfa einen Datenschnitt gemacht über die entsprechenden Bewertungsverfahren bis zum damaligen Zeitpunkt. Da fußten acht von 17 Bewertungen, wenn ich das richtig erinnere, auf einer einzigen Studie. Das heißt, das ist ja kein so seltenes Phänomen. Nun kann man sich dessen entledigen, indem man sagt: Na und, wie machen hier einen Hypothesenshift. Wir treffen Annahmen. Wir wollen das. Wir müssen uns so positionieren. Das ist unser gesetzlicher Auftrag. Aber ich glaube, das wird nicht der Realität gerecht. Das ist eine weitere Herausforderung auch an die Methodiker, so gut es geht ihre Gegebenheiten in den Vorschlägen mit einzubeziehen. Sonst laufen wir in die Irre, weil wir können nicht gleichzeitig zwei Herren dienen. Das ist sehr schwierig. Das kann ich Ihnen garantieren. Ich bin Doppelstaatler. Und es ist für mich sehr schwierig, zwei Ländern zu dienen.

Thomas Kaiser: Zunächst einmal ist es ja nicht verboten, zwei Studien durchzuführen. Selbst wenn die Zulassungsbehörde sagt, uns reicht eine Studie für die Zulassung, ist es ja nicht verboten, eine zweite Studie durchzuführen. Das könnten Sie ja machen. Insofern haben Sie

das als Hersteller schon in der Hand. Ich weiß, dass das mehr Aufwand ist, mehr kostet, aber ein bisschen den Satz von Guido Skipka aufgreifend „Nichts ist umsonst im Leben“: Sie haben es natürlich in der Hand. Wenn Sie das weiterdenken, anzupassen an diese eine Situation mit der einen Studie, dann muss man sich natürlich auch fragen: Was bedeutet das denn für eine Situation mit drei oder vier Studien? Das gibt es natürlich auch in Zulassungsverfahren. Wenn man an die Antidepressiva denkt, an unsere zurückliegenden Bewertungen, da gab es Situationen mit sechs oder sieben Studien. Damit würden Sie natürlich deutlich niedrigere Schwellenwerte ansetzen. Ich glaube, wenn man versucht, jeglichen äußeren Einfluss irgendwie berücksichtigen zu wollen, wird es relativ beliebig in der Festlegung von solchen Operationalisierungen für Schwellenwerte. Insofern denke ich, dass wir mit der jetzigen Vorgehensweise den Regelfall gut abgebildet haben. Man hat es als Hersteller auch in der Hand, nicht nur für die Zulassung, sondern auch für die Nutzenbewertung zu arbeiten.

Guido Skipka: Ich möchte mal den wichtigsten Faktor nennen, der das Ausmaß bestimmt, das ist der Effekt, der tatsächlich vorliegt. Ob wir über eine oder zwei Studien reden, Fallzahlen hin und her, ein Ausmaß wird in erster Linie dadurch bestimmt, wie groß der Effekt ist. Das hat mit seltenen Erkrankungen oder sonst etwas wenig zu tun. Vielleicht muss man sich mal von da her dem ganzen Punkt nähern. Wenn eine Therapie einen großen Effekt hat, wird man das auch mit einer Studie feststellen. Das sollte man vielleicht nicht vergessen.

Stefan Lange: Ich wollte ergänzen, dass wir bitte möglichst nicht Zulassungsfragen mit unseren Problemen vermengen. Wenn in der Zulassung eine Studie ausreicht, was ja in der Regel nicht der Fall ist, weil Sie noch andere Untersuchungen dazu haben, aber eben eine große klinische Studie ausreicht, dann ist das eine Angelegenheit der Zulassung. Ob und wie das für die Nutzenbewertung eine Rolle spielt und was das für die Nutzenbewertung bedeutet, ist eine ganz andere Frage. Insofern hat Thomas Kaiser natürlich völlig Recht, dass die Zulassung meines Wissens bisher nicht verboten hat, Studien durchzuführen. Dann kenne ich jedenfalls noch nicht, wäre neu. Das wäre interessant zu erfahren.

Ich wollte nur auf einen Punkt hinweisen, der ein bisschen das ergänzt, was Guido Skipka gesagt hat. Wir haben ja zahlreiche Beispiele, Herr Leverkus - ich habe mal eine kleine Sammlung angefangen; die könnte ich Ihnen mal schicken -, wo bei entsprechenden Effekten auch in einer Studie, in kleinen Studien all diese Schwellenwerte erreicht werden können. Es hängt eben davon ab, was in einer Studie gezeigt worden ist, was als Effekt da ist. Noch einmal: Es gibt zahlreiche. Das wissen wir aus der Erfahrung aus den früheren Nutzenbewertungen. Sie kennen ja auch die Verteilung der Ergebnisse. Es ist ja nicht so, als ob wir in zwei Drittel der Fälle keinen Zusatznutzen hätten, in den weiteren fünf Sechstel von dem verbleibenden Drittel maximal einen geringen Zusatznutzen und vielleicht in einer Bewertung einen beträchtlichen. Schauen Sie sich mal die Verteilung an. Das Ganze funktioniert doch im Grunde genommen, wenn man ehrlich ist, ganz gut. Die entscheidende Bitte: Bitte nicht Zulassungsaspekte vermengen mit den Fragen, die wir bei der frühen Nutzenbewertung haben.

Moderator Jürgen Windeler: Jetzt bitte ich Herrn Schinzel, vielleicht mit dem Hinweis, dass wir ein bisschen aufpassen, uns nicht in breiten Diskussionen über die Ausmaßfestsetzung als solche zu verlieren. Das ist bestimmt spannend, aber jetzt hier eigentlich nicht primäres Thema.

Stefan Schinzel: Ich möchte Herrn Skipka eigentlich widersprechen, und zwar aus dem Beispiel, das wir hatten, mit unserem Dossier zu Cabazitaxel. Wir haben keine Halbierung des Mortalitätsrisikos gehabt, sondern eine Reduktion um 30 %, aber wir hatten halt eine Fallzahl, die für den signifikanten Nachweis viel zu groß war. Das war im Grunde genommen aus Unwissenheit in der Planung heraus. Wir hätten insgesamt, wenn Sie das Gesamtergebnis der Studie anerkannt hätten, durchaus einen erheblichen Zusatznutzen gehabt. Dass es dazu nicht gekommen ist, lag an der Aufteilung gemäß Alterskategorien.

Und der zweite Punkt, den ich gerne ansprechen möchte, ist diese Annahme eines erheblichen Zusatznutzens bei der Mortalität, dass da eine Halbierung der Mortalität vorliegen sollte: Dieses System, das Sie benutzen, an Punktschätzern, um im Grunde genommen diese Schwellenwerte für die Obergrenze der Konfidenzintervalle abzuleiten - das ist ja im Grunde genommen eine Herleitung gewesen, die in den Simulationen, die Sie in Ihrem Papier präsentieren, insbesondere für die niedrigen relativen Risiken -, spiegelt ja nicht die Annahmen wider. Sie bekommen ja in diesen Simulationen im Grunde genommen Werte für die relativen Risiken, die deutlich über denen liegen, die Sie als Annahme investiert haben in Ihrer Herleitung. Die Frage ist, ob das aus Ihrer Sicht ein separater Punkt ist, aber ich denke, das sollten wir auch noch mal andiskutieren.

Stefan Lange: Habe ich Sie richtig verstanden, dass Sie unglücklich damit sind, dass wir Ihr Medikament so positiv bewertet haben?

Stefan Schinzel: Sie haben es ja nicht so positiv bewertet, weil sie den Split nach Alter gemacht haben. Aber wenn man das Gesamtergebnis anerkannt hätte, hätte man festgestellt, dass im Grunde genommen die Mortalitätsreduktion 30 % beträgt, es aber trotzdem für einen erheblichen Zusatznutzen reicht, weil wir eben eine sehr viel größere Fallzahl gehabt haben, als wir für den signifikanten Nachweis gebraucht hätten.

Guido Skipka: Das ist ein separater Punkt. Den sollten wir vielleicht später diskutieren. Ich muss auch ehrlich gesagt darüber nachdenken. Ich bin nicht sicher, ob ich Sie richtig verstanden habe, weil die Grenzen, die jetzt durch die Simulation zustande kamen… Es ging ja, um es ganz kurz zu erklären - in der Rationale ist es ja dargelegt; die asymptotische Formel, die wir verwendet haben, die hat sich als nicht besonders gut herausgestellt; deswegen haben wir diese Simulation nachgeschoben -, in meinen Augen eher darum, zu bestätigen, dass wir eigentlich keinen großen Fehler da begehen.

Moderator Jürgen Windeler: Letzte Bemerkung dazu. Ich möchte den BPI schon mal vorwarnen, dass ich Sie gleich wieder zu dem Punkt von vorhin frage.

Ch.-Markos Dintsios: Ich greife noch einmal das auf, was Herr Schinzel gesagt hat. Das könnte auch zu einer falschen Anreizsetzung führen. Denn in diesem Beispiel hätte man das jetzt nicht aufgeteilt aufgrund des Effektmodifikators Alter - das sind zwei Subgruppen -, übrigens Kommentar Stefan Lange. Wenn man sich das auf Subgruppenebene anschaut, schaut das Ergebnis der frühen Nutzenbewertung ein bisschen anders aus. Da ist man nicht mehr bei zwei Drittel positiv, sondern hälftig, und wenn man das auf einen Prävalenzansatz macht, wird das nur bei einem Fünftel positiv sein. Das ist eine Betrachtungssache, wie man sich das Ganze maßschneidert.

Ich komme zurück zu meinem ursprünglichen Punkt. Der falsche Anreiz, der damit gesetzt werden könnte, ist, dass man - ganz einfach - in der Annahme, man will einen Effektschätzer erreichen, überpowert. Das geht auch. Das wird ein bisschen teurer. Denn dann haben Sie in meinen Augen eine kontraintuitive Reaktion geschaffen. Sie haben für die Herleitung der oberen Schwellen in Ihrem Hypothesenshiftansatz eigentlich Effekte eingebaut. Sie haben sich schon Djulbegovic‘schen Anker festgemacht, für Sie - wie gesagt, freundlich zu verstehen, nicht sehr höflich -, mit dem 0,5er Hazard Ratio, setzen das über alle relativen Effektmaße gleich, Odds Ratio relative Risiken, das hat, glaube ich, auch einige Fragen aufgeworfen, und kommen zu dem Ergebnis, nach der Simulation mit den oberen Konfidenzintervallen, machen sozusagen die Klassifizierung des Ausmaßes an der Präzision fest. Und die Frage ist: Würde das nicht dazu führen können - es ist alles hypothetisch, was wir hier diskutieren; wir sind ja eigentlich in einem geschützten Rahmen -, dass jemand, der auf einen erwarteten kleinen Effektschätzer, als Beispiel eine Mortalität von 0,5 - hier war das Beispiel 0,3; es kann auch bei 0,28 liegen -, so stark powert, dass er eben diese obere Konfidenzintervallgrenze immer noch unterschreitet, die Sie über den Hypothesenshift hergeleitet haben. Die Warnung ist: Man kann durchaus auch falsche Anreize setzen mit solchen Vorgehen.

Moderator Jürgen Windeler: Ich möchte bemerken wollen, dass das nicht hypothetisch ist, was wir hier machen, sondern die konkrete Frage, die wir gerade versuchen zu klären, ist: Ist es sinnvoll, die Konfidenzintervallgrenzen zu benutzen oder die Punktschätzer? Bisher schienen mir die Äußerungen in diese Richtung ziemlich eindeutig zu sein.

Ich habe jetzt noch drei Wortmeldungen, dann würde ich diese Diskussion beenden wollen. Ich bitte auch, bei diesem Thema zu bleiben.

Dieter Hauschke: Vielleicht noch einmal zu dem Punkt Effektschätzer oder Effekt. Man muss da ein bisschen differenzieren. Man kann nicht einen Effektschätzer overpowern. Der Effektschätzer hängt von dem unbekannten Effekt ab. Da vermengen Sie ein bisschen die Methodik. Es ist wirklich so: Die Studie hängt eigentlich davon ab, wie gut das Präparat ist. Das ist der unbekannte Effekt. Wenn Sie einen schlechten Effekt haben, können Sie nach wie vor so powern bzw. auf den Punktschätzer schauen. Man sollte das also ein bisschen differenziert betrachten. Es ist richtig, Konfidenzintervalle zu verwenden, in diesem Fall die obere. Man kann darüber diskutieren, wie die Grenzen sind, aber die statistische Methode basiert nun einmal auf Konfidenzintervallen und nicht auf Effektschätzern.

Ralf Bender: Kurze Bemerkung zu Markos: Er hat gesagt, wir würden die Schwellen für alle Effektmaße verwenden, Odds Ratio, Risk Ratio usw. Das ist einfach verkehrt. Ein Blick in den Entwurf des Methodenpapiers gibt Aufklärung.

Das zweite, was ich sagen wollte, hat Dieter Hauschke schon gesagt. Das hat sich damit erledigt.

Stefan Schinzel: Auch auf die Gefahr hin, dass ich mich wiederhole: Wenn Sie zwei Präparate haben, die die Mortalität um 30 % reduzieren, und bei dem einen Präparat treiben Sie die Fallzahl sehr hoch, dann können Sie die Obergrenze des Konfidenzintervalls näher an 0,7 heranbringen und unterschreiten damit eine Schwelle, die Sie für einen erheblichen Zusatznutzen brauchen. Insofern bitte ich um Nachsicht, aber ich verbleibe bei meiner

Position, dass hier eine Vermengung des Therapieeffekts mit der Präzision seiner Bestimmung stattfindet.

Moderator Jürgen Windeler: Ich schlage vor, dass wir das jetzt so stehen lassen, insbesondere die Positionen. - Jetzt frage ich die beiden Vertreter von BPI: Sind Sie sprachfähig? Die Frage haben Sie noch im Kopf? Dann bitte ich Sie jetzt um Ihre Erläuterung.

Ilona Krug: Die Frage habe ich insoweit im Kopf, als es um die festen Konfidenzintervallgrenzen ging. Und unsere Stellungnahme dazu war, dass bestimmte Populationen durch die Verwendung der festen Konfidenzintervallgrenzen benachteiligt sind. Unsere Stellungnahme dazu ist, dass wir der Meinung sind, dass eben Studiensettings mit großem Vorwissen benachteiligt sind, weil der Behandlungseffekt eben bereits gut eingrenzbar ist. Da ist praktisch eine genaue Studienplanung möglich, um halt zu einem signifikanten Ergebnis zu kommen. Bei diesen Studien liegt dann das Konfidenzintervall vermutlich knapp um den Nullpunkt. Es wäre halt unethisch, die Studie größer zu planen als notwendig. Von daher sagen wir, dass es starre Grenzen sind und dass es, egal um welche Populationen es sich handelt oder die Größe der Populationen, zu starre Grenzen sind.

Guido Skipka: Ich muss noch mal zurückfragen. Sie verwirren mich jetzt mit der starren Grenze. Das Argument, so wie ich es verstehe, ist: Sie planen eine Studie, Sie haben großes Vorwissen, können also sehr präzise planen - vielleicht kommt das auch so präzise heraus, wie man das so plant -, und man bekommt ein Konfidenzintervall, das gerade den Nulleffekt ausschließt. Das ist klar, dass man damit dann keine Schwelle, wie auch immer sie geartet ist, erreichen kann, die irgendwie entfernt vom Nulleffekt ist. Das hat mit einer starren Grenze aber meiner Meinung wenig zu tun. Das wird mit jeder Schwelle bei jeder Grenze scheitern. Das ist aber nicht unsere zentrale Frage gewesen. Die zentrale Frage - ich kann sie noch einmal wiederholen - ist: Wir bewegen uns hier in der frühen Nutzenbewertung. Dieses „früh“ besagt, dass wir in der Regel eben nicht diese umfangreichen Vorinformationen haben. Das ist das eine.

Das andere ist, dass diese Studien, die uns da vorgelegt werden - das sind Zulassungsstudien -, auf einen primären Endpunkt gepowert sind, der uns zum Teil überhaupt nicht interessiert, weil er nicht patientenrelevant ist. Das heißt, wir schauen auf die anderen Endpunkte, die in der Studie mit erhoben worden sind, hoffentlich patientenrelevante. Für die kann das Argument ja gar nicht zutreffen, weil man ja gar nicht auf diese Nebenendpunkte hin gepowert hat.

Moderator Jürgen Windeler: Möchten Sie das noch erläutern?

Ilona Krug: Nein, das nehme ich so zur Kenntnis, danke.

Moderator Jürgen Windeler: Wenn es keine weiteren Wortmeldungen gibt, dann würde ich diesen TOP nun abschließen.

Wir kommen zum

Tagesordnungspunkt 2:

Verwendung von Ereignissen oder Nichtereignissen für das relative Risiko

Guido Skipka: Wir haben im Institut festgelegt, wenn wir es mit binären Daten zu tun haben, sprich mit der Vierfeldertafel, dass wir das relative Risiko betrachten. Wir haben lange überlegt, ob man vielleicht auch Odds Ratio oder noch etwas anderes nehmen kann. Wir sind letztendlich zum relativen Risiko gekommen aus der schon vorhin angesprochenen Betrachtungsweise, dass dann, wenn man zwei Studien - entsprechend gepowert - hat, die Schwellenwerte so aussehen, dass man dann auch die entsprechende Power hat. Wenn man die asymptotische Formel, die wir seinerzeit dazu verwendet haben, die in der SAS-Prozedur abgelegt ist, verwendet, stellt man fest, dass im Gegensatz zum Odds Ratio beim relativen Risiko die schöne Eigenschaft herauskommt, dass die Schwellenwerte unabhängig vom Basisrisiko sind. Das ist natürlich erst einmal eine wunderbare Eigenschaft, weil man nicht auch noch in jedem Setting irgendwelche Basisrisiken ermitteln und dann noch die Schwellenwerte danach adjustieren muss. Das hatte sich nach den Formeln so herausgestellt. Das war unser Hauptbeweggrund für das relative Risiko. Wer die Rationale gesehen hat, hat ja gesehen, die Formel hat gewisse Schwächen, sprich die Asymptotik funktioniert nicht in allen Situationen so ganz gut. Deswegen haben wir das ein bisschen nachuntersucht. Aus unserer Sicht ist das nicht besonders kritisch. Das ist aber nur Vorgeplänkel.

Das eigentliche Argument, das ich hier diskutieren möchte, ist: Wenn man das relative Risiko nimmt, hat man eigentlich noch ein Problem, weil es sich genau genommen um zwei Maße handelt: Es hängt nämlich von der Blickrichtung ab, von der ich draufschaue. Betrachte ich Ereignisse oder Nichtereignisse? Kurzes Zahlenbeispiel: Wenn ich in der einen Gruppe 10 % Risiko habe, in der anderen 20 %, komme ich zu einem relativen Risiko von 0,5. Betrachte ich die Gegenereignisse, also 80 % zu 90 %, dann liege ich irgendwo in der Nähe von 0,9. Wenn man sich also auf das relative Risiko einigt, muss man schauen, gucke ich von der rechten oder von der linken Seite, wenn ich das mal so salopp ausdrücken darf.

Wir haben im Methodenentwurf dazu geschrieben, dass man in der konkreten Situation anhand inhaltlicher Kriterien eben festlegen muss, von welcher Seite man das relative Risiko betrachtet. Das ist in der Stellungnahme von Sanofi-Aventis noch einmal aufgegriffen worden. Es wird kritisiert, dass wir das nicht genauer operationalisieren, wann wir von welcher Seite auf das relative Risiko schauen. Damit sei die Tür geöffnet für eine ergebnisgetriebene Festlegung. Sanofi-Aventis sagt richtigerweise, durch die Verwendung des Odds Ratios als Abstandsmaß hätte man dieses Problem erst gar nicht gehabt. Das ist sicherlich richtig. Ich habe Ihnen gerade erläutert, wie wir zum relativen Risiko kamen. Wir könnten es uns jetzt vielleicht ganz einfach machen und sagen: Wunderbar, dann nehmen wir halt das Odds Ratio. - Das hat natürlich auch Implikationen auf die Ableitung der Schwellenwerte. Das muss man sich mal genau überlegen. Meine eigentliche Frage ist aber, ob Sie, wenn wir so tun, als wäre das relative Risiko ein geeignetes Maß, Vorschläge hätten, wie man die Wahl der Blickrichtung operationalisieren könnte.

Stefan Schinzel: Ich denke, in vielen Situationen ist es eigentlich naheliegend. Wenn Sie zum Beispiel Patienten mit unerwünschten Ereignissen betrachten, dann finde ich es etwas

künstlich, etwas artifiziell, sich auch noch Patienten anzuschauen, die das entsprechende Ereignis nicht hatten. Leider sind Sie bei dem Aflibercept Dossier so verfahren, dass Sie bei unerwünschten Ereignissen, die zum Therapieabbruch geführt haben, sich beide Seiten angeschaut haben.

Was den zusätzlichen Schaden anbelangt, geben Sie dann eine Einschätzung von, bis. Die umfasst zwei benachbarte Kategorien, ich meine „beträchtlich“ und „erheblich“, wenn ich es richtig erinnere. Wir haben uns damals ein bisschen mit diesem Szenario beschäftigt. Und wir haben darüber hinaus auch Situationen gefunden, wo sogar eine Klasse übersprungen wird, also „gering“ und „erheblich“ oder „kein“ und „beträchtlich“. Wir warnen eigentlich davor, dass hier im Grunde genommen bei unerwünschten Ereignisse, die zum Abbruch führten, das von beiden Seiten betrachtet wird. Da waren wir eigentlich ziemlich unglücklich. Das muss ich ganz ehrlich sagen. Das wäre eine Situation, denke ich,… Aus meiner Sicht gäbe es da nichts zu diskutieren. Es gibt sicherlich klinische Endpunkte. Wir sind gerade dabei, ein Dossier fertigzustellen. Da geht es um die Frage: Hat ein Patient mit Multipler Sklerose im bestimmten Zeitraum einen Schub oder nicht? Da ist die Sache sicherlich tüfteliger oder schwieriger. Aber in dem genannten Beispiel zu Aflibercept habe ich nicht verstanden, warum dieser Übergang zum Gegenereignis bei Patienten ohne unerwünschte Ereignisse stattgefunden hat. Der hat uns im Grunde genommen nur eine Verschlechterung des zusätzlichen Schadens eingebracht. Darüber haben wir uns auch, wenn ich das so offen sagen darf, ein bisschen geärgert.

Moderator Jürgen Windeler: Danke für den Hinweis.

Thomas Kaiser: Wir haben das ja in der Anhörung zu dem Bericht beim Gemeinsamen Bundesausschuss diskutiert. Ich kann Ihre Anmerkungen an der Stelle auch erst einmal verstehen. Nur an der Gesamtbewertung - das habe ich in der Anhörung gesagt, und das habe ich Ihnen auch geantwortet; ich weiß nicht, ob ich Ihnen persönlich oder Ihrer Firma auf das Schreiben, das Sie geschickt haben, geantwortet habe - hat das überhaupt nichts geändert. Das heißt, am Ende steht ja noch die Gesamtabwägung. Und die ist durch dieses Vorgehen „von, bis“ überhaupt nicht beeinflusst worden. Nichtsdestotrotz haben wir diesen Punkt aufgenommen und können dem auch folgen, dass es im Grunde genommen eine Art natürliche Richtung für bestimmte Dinge gibt. So ist das im Methodenpapier ja auch formuliert. Die Aflibercept-Bewertung als solche bzw. die Erarbeitung dieses Punktes war noch genau in diesem Zeitraum, als wir hier diesen Methodenanhang zur Stellungnahme erarbeitet haben.

Stephanie Rosenfeld: Zur Ergänzung: Sie sagen - wir haben ja auch einen Brief von Ihnen bekommen -, es hätte letztendlich an der Bewertung nichts geändert. Es waren natürlich auch noch andere Punkte, die dann natürlich schon etwas an der Gesamtbewertung geändert hätten. Wir wollten jetzt nicht noch einmal einen Antwortbrief schreiben. Letztendlich war das mit einer der beiden strittigen Punkte. Zu dem anderen Punkt haben Sie sich in dem Brief nicht geäußert. Das ist jetzt Nutzenbewertung speziell ein Produkt.

Ich wollte noch ergänzen: Sie machen ja auch den Vorschlag, vorher festzulegen, von welcher Seite das Ereignis betrachtet wird. Meine Frage ist im Rahmen der Nutzenbewertung: Wo hat der Hersteller das Portal? Woher weiß also der Hersteller, von welcher Seite das IQWiG

herangeht? Wir erfahren das in der Regel ja erst zu dem Zeitpunkt, wo die IQWiG-Nutzenbewertung vorliegt.

Stefan Lange: Ich glaube, man sollte versuchen, sich von konkreten Beispielen oder Ärgernissen zu lösen, so verständlich sie auch sein mögen. Wir meinen das schon sehr ehrlich, dass wir schon selber ein gewisses Problem sehen, wo wir Sie alle oder auch die, die bei Ihnen im Hintergrund sind, um vernünftige Vorschläge bitten. Wir haben uns innerhalb des Instituts lange mit der Frage beschäftigt: Gibt es so etwas wie eine natürliche Blickrichtung? Ich persönlich bin davon immer noch nicht überzeugt, muss ich ganz offen gestehen. Daran erkennen Sie auch, dass wir im Institut nicht immer eine einheitliche Blockmeinung haben, sondern durchaus hier diskutieren. Ich glaube, man kann sich verschiedene Dinge vorstellen. Aber bevor wir wieder unsere eigenen Ideen dazu präsentieren, wäre es ganz nett, wenn auch jemand von Ihnen mal eine Idee präsentieren könnte. Natürlich können Sie im Dossier zum Beispiel mit vernünftigen Argumenten selber die Blickrichtung vorgeben. Das wäre eine Möglichkeit. Ob das jetzt die Beste ist… Zum Beispiel: Ich wüsste gar nicht, was die natürliche Blickrichtung bei der Mortalität ist. Ob von links oder rechts die Menschen sterben, ist meistens egal, weil sie dann tot sind. Ich bin also nicht so überzeugt, ob das wirklich zielführend ist. Es ist vermutlich gar nicht so oft ein Problem; das mag wohl der Fall sein. In der Vergangenheit hat sich das für unsere Gesamtbewertungen nicht als Problem herausgestellt. Aber es kann tatsächlich mal auch für die Gesamtbewertung ein Problem sein. Und da wäre es toll, wenn es ein gemeinsames Verständnis von gutwilligen Methodikern gäbe, wie man mit dieser Situation umgeht. Das Odds Ratio ist eine Möglichkeit, weil es eben symmetrisch ist. Das relative Risiko ist nicht symmetrisch. Erstaunlich finde ich in diesem Zusammenhang, dass es praktisch kaum Publikationen gibt, dass es ein generelles Problem des relativen Risikos ist, wenn man sich über die Quantität des Effektmaßes oder des Effektschätzers Gedanken macht. Dieses Problem ist praktisch nicht betrachtet. Man könnte ja auch das relative Risiko symmetrisieren. Vielleicht hat das auch ganz interessante Eigenschaften. Ich weiß es nicht.

Ch.-Markos Dintsios: Stefan Lange hat ja aufgefordert, Vorschläge zu machen. Ich möchte keinen Vorschlag machen, aber mich würde interessieren, was in einem ähnlich gearteten Beispiel passiert, das so aussehen würde wie das, was vorhin diskutiert wurde, wo es in Abhängigkeit von der Blickrichtung zu unterschiedlichen Klassifizierungen kommen würde, sozusagen eine Ungleichgerichtetheit des Ausmaßes, ob man sich dessen auch bewusst sein muss.

Und vielleicht als Hilfeleistung oder Handlungsempfehlung: Weil es sich zumindest in der frühen Nutzenbewertung vornehmlich um Studien aus der Zulassung handelt, könnte man sich doch anschauen, wie die Zulassung damit umgegangen ist, in welche Richtung sie wollte. Ich bin Anhänger der Hypothesenlogik und würde sagen: Ich teste ja auch eine Hypothese und schaue mir die Ereignisse an. Vielleicht muss ich es nicht festmachen an den Gegenereignissen. Aber das wäre eine Möglichkeit. Das wäre offen. Das ist ein ad-hoc-Vorschlag, wie man mit so etwas umgehen könnte. Aber ich löse damit nicht das Problem, wenn ich mit dem relativen Risiko arbeite. Was passiert in der Gegenereignisbetrachtung, wenn es dann zu einer ganz anderen Klassifizierung kommen würde? In dem Beispiel wurde ja von, bis. Wir haben vorhin gehört, es gibt Konstellationen, wo man sogar drei Kategorien abhandeln kann bzw. eine überspringt und ein Teil inkludiert. Das ist nicht unbedingt auch für das IQWiG zielführend, mit solchen breiten Aussagen operieren zu müssen.

Thomas Kaiser: Man könnte sogar „im Gegenteil“ sagen, das aufgreifend, wenn man das mit „von, bis“ beschreibt für die einzelnen Endpunkte. Noch einmal: Das ist ja nicht die Gesamtaussage zum Zusatznutzen, sondern das ist auf eine Endpunktebene, wenn man das mit „von, bis“ beschreiben würde und hinterher darstellen würde, was das für die Gesamtaussage bedeutet. Hier kann ich nur noch mal sagen: Diese konkrete Endpunktbeschreibung hat an der Gesamtaussage nichts geändert. Es hat keine Bedeutung für die Gesamtaussage gehabt. Dann könnte man sagen: Das Problem des relativen Risikos spielt keine Rolle für die Gesamtaussage. - Dann kommt aber die Firma und sagt, sie ärgern sich darüber, dass das hier gemacht wurde, weil für einen Endpunkt dann „erheblich“ statt „beträchtlich“ steht. Wie man es also macht, macht man es falsch. Die einen sagen: Es gibt eine natürlich Richtung. Man sollte sich das gar nicht anders angucken. - Der nächste sagt: Man kann sich beide Situationen anschauen und dann gucken, ob das irgendeine Bedeutung hat. - Ich glaube, den Vorschlag von Stefan Lange aufgreifend - das gilt ja für viele Dinge im Dossier, auch für Abweichungen von der Zulassung etc. -: Im Dossier begründen, warum man auf eine bestimmte Art und Weise einen Endpunkt so oder so wertet, ein Ereignis oder ein Gegenereignis betrachtet, das macht sicherlich sehr viel Sinn. Denn der Hersteller ist in diesem Verfahren ja aufgefordert, die Evidenz beizubringen, die Grundlage für die Bewertung zu liefern und damit auch die Argumentation für die Bewertung. Dem verschließen wir uns nicht, sie muss nur vernünftig sein.

Guido Skipka: Markos, dein Vorschlag, zu gucken, welche Blickrichtung die Zulassungsbehörden einnehmen, klingt erst einmal gut, funktioniert nur leider nicht. Das erklärt vielleicht auch, was Stefan Lange sagt, warum man dazu so wenig Literatur findet. Die Art der Blickrichtung wird erst dann interessant, wenn man die Hypothesengrenze von Null verschiebt. Bei der zentralen Hypothese ist das unerheblich, ob man von rechts oder links draufschaut. Den p-Wert beeindruckt das nicht. Die Zulassungsbehörden haben - nageln Sie mich nicht fest - meiner Meinung nach wenig mit verschobenen Hypothesen zu tun, insbesondere für die Endpunkte, die wir betrachten. Von daher werden wir da in den Zulassungsunterlagen keine Hilfestellung finden.

Stefan Schinzel: Zunächst möchte ich kurz Stellung nehmen zu der Einlassung von Herrn Dr. Kaiser. Es ist so: Bei einem anderen unerwünschten Ereignis, bei einem anderen Typ hätte man durch Betrachtung des Gegenereignisses im Grunde genommen eine Stufe herunterkommen können beim zusätzlichen Schaden. Das haben Sie nicht getan. Sie haben im Dossier nur den Fall ausgearbeitet, wo es zu einer Verschlechterung, zum Nachteil eines Präparates geführt hat. Das fehlt vielleicht, um zu erläutern, warum ich mich geärgert habe.

Der zweite Punkt ist: Mir ist nicht ganz klar, warum man nicht doch etwas ernsthafter das Odds Ratio in Betracht ziehen will. Wenn die Risiken gering sind, ist das Odds Ratio in derselben Größenordnung wie das relative Risiko. Wenn fast alle Patienten ein unerwünschtes Ereignis haben, dann entspricht das Odds Ratio der relativen Chance, ereignisfrei zu bleiben. Nur in diesem Mittelbereich hat man halt eine stärkere Abweichung. Aber insgesamt gesehen hat das Odds Ratio doch sehr viel schönere symmetrische Eigenschaften.

Ich verstehe natürlich, dass man nach 50 Nutzenbewertungen in der Zwangslage ist. Das ist die normative Kraft des Faktischen. Aber warum ist das kein Diskussionspunkt, ob man nicht vielleicht doch Odds Ratios angucken sollte?

Stefan Lange: Ich glaube, das ist ein Missverständnis. Den Druck der normativen Kraft des Faktischen verspüren wir nicht. Noch einmal: Wenn wir gute Vorschläge bekommen, wie wir ein konkretes Problem beseitigen können, sehen wir uns dazu durchaus in der Lage. Sprich: Natürlich können wir uns vorstellen, das Effektmaß zu wechseln. Warum nicht? Aber das hat natürlich, finde ich, ein ganz spezifischen Problem neben dem, was Guido Skipka geschildert hat, neben der zunächst einmal schönen Eigenschaft. Im Zusammenhang mit der Herleitung unserer Schwellenwerte haben wir es ja häufig mit time-to-event-Analysen zu tun. Da liegt natürlich das Hazard Ratio als Effektmaß dem relativen Risiko näher als das Odds Ratio. In diesen Situationen hätten wir ganz spezifische Probleme, wie wir damit umgehen. Insofern klammern wir uns ein bisschen an dem relativen Risiko. Das ist ein Grund dafür. Wie gesagt, ich könnte mir auch andere Ansätze vorstellen. Ich weiß nur nicht, wie sie von ihren statistischen Eigenschaften her aussehen. Ich habe bereits ein Stichwort genannt: Man kann ja das relative Risiko im Grunde genommen symmetrisieren. Aber, wie gesagt, da kenne ich die Eigenschaften nicht. Das wäre eine Aufgabe für Methodiker, Statistiker, sich das einmal anzuschauen.

Zur Einlassung von Markos Dintsios: Dieses Problem wird schlicht und ergreifend nicht adressiert bei Zulassungsbehörden. Das interessiert sie nicht. Ich glaube schon, dass auch in der Zulassung verschobene Hypothesen - Nichtunterlegenheit ist ein Stichwort - eine Rolle spielen, und könnte mir vorstellen, dass es da vielleicht auch mal ein Problem sein könnte. Ich glaube, es ist bisher als Problem nicht erkannt.

Stephanie Rosenfeld: Ich möchte gerne auf Herrn Kaiser eingehen. Er hat vorgeschlagen, der Hersteller sollte doch definieren, von welcher Seite das Ereignis betrachtet werden soll. Er wäre ja auch verpflichtet, Evidenz beizubringen. Da stimme ich Ihnen vollkommen zu. Auch der Hersteller ist wirklich interessiert, die Evidenz zu bringen. Nur in vielen Fällen ist es so, dass diese natürliche Blickrichtung nicht so gegeben ist. Ich denke, jede Firma wird bereit sein, auf bestimmte Wünsche einzugehen. Die Hersteller haben jedoch das Problem, zu erfahren, wie die IQWiG-Sicht ist. Der G-BA berät ja nicht zu Details der Studien oder zu Details im Dossier. Also: Wo haben wir den Austausch? Ich könnte mir gut vorstellen, dass ein Austausch im kleinen Kreis möglich ist. Das muss nicht immer im großen Rahmen mit G-BA und allem sein, sondern dass man vielleicht einen Ansprechpartner innerhalb der Biostatistik hat, wo man solche Dinge besprechen könnte.

Moderator Jürgen Windeler: Im Sinne einer für die Hersteller besser vorhersehbaren, kalkulierbaren Richtung könnte eine Symmetrisierung, wie Herr Lange es beschrieben hat, durchaus ein Weg sein, wo man sich um die Inhalte und um so genannte natürliche Blickrichtungen gar nicht mehr kümmern muss.

Thomas Kaiser: Ich denke auch, dass man deswegen entweder eine allgemeingültige Lösung oder eine Gesetzesänderung braucht. Dafür sind wir die falschen Ansprechpartner. Wenn Sie eine Beratung durch das IQWiG wollen, dann müssen Sie es an den Gesetzgeber geben, dass er im SGB V das beschreibt. Das kann man nicht anders sagen. Man kann nicht sagen: Ja klar, gerne, machen wir. - Das passt in diesem Verfahren so nicht.

Stephanie Rosenfeld: Das ist eben ein Stück weit das Problem, dass man halt immer erst hinterher die Möglichkeit zur Diskussion hat. Schöner wäre es vorher.

Thomas Kaiser: Noch einmal: Der Gesetzgeber steht dafür offen.

Moderator Jürgen Windeler: Vielleicht darf ich mir die Bemerkung erlauben: Der Gesetzgeber regelt die Beratung innerhalb konkreter Verfahren, dass man sich generell auf methodischer Ebene mit Fachgesellschaft und auch mit Industrievertretern zusammensetzen und sagen kann, was ein vernünftiges Verfahren ist, oder besser: Welche Vor- und Nachteile haben die einzelnen Dinge, die man sich vorstellen kann? Das ist eine andere Ebene.

Stefan Schinzel: Ich habe eine kurze Bemerkung: Diese Symmetrisierung würde natürlich erforderlich machen, völlig neu über Schwellenwerte nachzudenken, es sei denn, Sie nehmen das Maximum. Dann sind wir natürlich gekniffen, und Sie können die alten Regeln verwenden.

Stefan Lange: Ich kann direkt darauf antworten, obwohl ich glaube, dass diese Erörterung nicht das richtige Forum dafür ist. Ich bin sehr dafür, dass wir ein Forum dafür schaffen. Noch einmal: Meine herzliche Bitte an Sie alle oder die Menschen, die bei Ihnen im Hintergrund arbeiten, in der Statistik: Beschäftigen Sie sich mit dem Problem. - Dann machen wir einen kleinen Workshop, Tagung oder Symposium dazu und gucken, welche Vorschläge es gibt. Sie sagen, wir sind alle gekniffen, wenn man das Maximum nimmt. Beim Odds Ratio nimmt man auch das Maximum, weil es symmetrisch ist. Je nach Blickrichtung wiederum ist man vielleicht nicht gekniffen, sondern in der glücklichen Lage, weil wir ja immer zwei Alternativen miteinander vergleichen. Das Leid der einen Alternative ist die Freude der anderen. Ich glaube, man muss versuchen, sich vom Einzelfall zu lösen. Der Einzelfall ist immer schwierig. Das macht nicht viel Sinn. Ziel kann eigentlich nur sein, eine allgemeingültige Lösung zu finden. Letztes Mal der Appell: Versuchen Sie, uns gute Vorschläge zu machen, um gemeinsam Lösungen für dieses Problem zu erarbeiten. Da sind wir sehr offen.

Ch.-Markos Dintsios: Ich wollte noch auf etwas hinweisen. Natürlich läuft man hier Gefahr, wenn man beide Seiten simuliert und sich das Ergebnis anschaut, weil es dann fakultativ offengelassen wird mit einer entsprechenden Rationale, die allerdings im Dossier bedient werden muss, auch ergebnisgesteuert die Richtung im Nachgang festlegen zu wollen. Was mich als Verbandsvertreter an diesem Vorgehen ein bisschen stört, ist, dann hat sich einer, der sich anhand der Ergebnisse, die er simuliert hat, festgelegt, hat eine Rationale aufgebaut, und dann kommt der nächste, der gerne denselben Endpunkt im selben Indikationsgebiet aus der anderen Richtung bedient sehen haben. Das ist natürlich eine Art Präzedenzfall, wer als erster kommt, mahlt als erster. Das ist für mich ein gewisses Risiko. Ich kann die Argumente von Stefan Lange sehr gut nachvollziehen, damit Sie mich nicht falsch verstehen, aber es ist ein gewisses Risiko, die methodische Fragestellung ein bisschen zu politisieren. Ist daran im IQWiG schon einmal gedacht worden?

Moderator Jürgen Windeler: Ich glaube, ich habe das nicht falsch verstanden. Sie haben gesagt: Es macht Sinn, eine übergreifende abstrakte Lösung zu generieren und es nicht jedem einzelnen zu überlassen, sich selber zu überlegen, was er optimieren möchte. - Da sind wir, wie Stefan Lange vorhin gesagt hat, sowohl interessiert als auch gerne bereit, uns zu beteiligen. Insofern haben wir darüber nachgedacht, wie wir mehrfach gesagt haben. War das jetzt nicht das, was Sie gemeint haben?

Ch.-Markos Dintsios: Dann bin ich von der falschen Ausgangssituation ausgegangen. Es war doch so, dass angedacht war, es fakultativ offen zu lassen und eine Rationale dafür anzubieten, aus welcher Richtung man kommt. Korrigieren Sie mich, wenn ich es falsch verstanden habe, Herr Windeler.

Moderator Jürgen Windeler: Unser Appell immer wieder und sicherlich nicht zum letzten Mal heute ist, mit konkreten Vorschlägen auf Probleme, die wir mit eigenen Vorschlägen zu lösen versuchen, zu reagieren. Der Vorschlag von Stefan Lange war: Eine Überlegung könnte sein, jeder begründet in dem Dossier selber, wie er das macht. Das war aber die Situation, die schon von anderen als nicht ganz befriedigend beschrieben worden ist. Aber die erstbeste Lösung ist es erst einmal. Jedenfalls haben wir dann eine Grundlage und einen Grund dafür. Eine eher abstraktere, übergreifendere Lösung macht sicherlich sehr viel mehr Sinn, wie Sie es auch gesagt haben, damit eben nicht einzelne möglicherweise auch sehr individuell zu Lösungen kommen, die für andere dann nicht mehr optimal sind. Es macht also, glaube ich, Sinn, das Problem als solches, was ein gravierendes Problem für die Abwägung zwischen Nutzen und Schaden ist… Denn wir gewichten in dieser Vorgehensweise, die wir im Moment machen, diese beiden unterschiedlich. Das ist auch der Grund gewesen, warum wir uns in verschiedenen Situationen überhaupt mit der Symmetrisierung beschäftigt haben. Das ist ein gravierendes Problem. Insofern sind wir daran interessiert, nicht an einer individuellen Lösung - solange wir die nicht haben, ist es aber trotzdem gut; immer noch besser als gar keine Lösung -, sondern an einer allgemeinen Lösung zu arbeiten. Wir wünschen uns und appellieren an Sie, sich an solch einer Erörterung zu beteiligen.

Dieter Hauschke: Ich weise darauf hin: Wir haben von der Biometrischen Fachgesellschaft und von der GMDS eine Präsidiumskommission IQWiG. Wir haben wiederholt aufgefordert, uns methodische Problemstellungen zu nennen, aber die Resonanz war relativ dürftig. Ich finde auch, immer zu den Methodikern abzuschieben… Es ist der originäre Anspruch der Pharmafirma, ihre Beiträge selbst zu bringen. Dadurch kann man auch einiges steuern.

Aber noch einmal: Sie sollten sich dann bitte bei den Fachgesellschaften melden. Und wir können darüber mit dem IQWiG diskutieren, was wir auch schon oftmals gemacht haben.

Guido Skipka: Ich bin der Meinung, wir müssen uns mal einigen, was wir eigentlich wollen. Mir ist das jetzt aufgefallen bei deiner Äußerung, Markos. Wir sind ja - das sieht man dem Entwurf an - bemüht, möglichst zu operationalisieren, zu abstrahieren, um diese Diskussion nicht zu haben, um die ergebnisgesteuerten Dinge aus den Weg zu schaffen. Das haben wir durch die festen Schwellenwerte versucht zu erreichen.

Wenn man sich die Stellungnahmen anguckt: Wir werden nahezu geprügelt dafür. Es wird gesagt: Das kann man doch nicht machen. Das muss doch indikationsspezifisch sein. Das hängt von a, b, c, d, e, f, g ab, und das muss man doch in der Indikation betrachten. Wenn wir an dieser Stelle im Entwurf schreiben: „Man muss im Einzelfall nach inhaltlichen Kriterien schauen, ob man von rechts oder links auf das relative Risiko schaut“, dann wird uns gesagt: Da müssen wir doch operationalisieren können. Das geht doch gar nicht anders. - Mein konstruktiver Appell ist: Wir müssen uns einigen: Wollen wir eine operationalisierte Lösung, die dann auch frei von Diskussionen ist, oder wollen wir in jedem individuellen Fall alles neu aushandeln, oder wollen wir irgendetwas dazwischen?

Moderator Jürgen Windeler: Das können wir als Schlusswort nehmen und auch für die folgenden Diskussionspunkte weiter im Sinn behalten.

Wir kommen zum

Tagesordnungspunkt 3:

Kriterien für Belege aus einer Studie

Stefan Lange: Wir haben dieses Thema schon kurz gestriffen. Das hat ein bisschen mit den ersten Diskussionen zu tun. Es ist schon angesprochen worden, dass wir neben der Feststellung des Ausmaßes, also der quantitativen Größenordnung, auch eine qualitative Feststellung machen, nämlich über die Beweislage, also wie stark die Evidenz ist. Ist es tatsächlich ein Beweis oder, wie wir früher gesagt haben und immer noch sagen, nur ein Hinweis oder neuerdings - eine etwas schwächere Kategorie - ein Anhaltspunkt? Das ist ja auch in der Arzneimittelnutzenverordnung so angelegt. Die Wahrscheinlichkeit für das Ausmaß des Zusatznutzens soll ja auch beschrieben werden, also nicht nur das Ausmaß, sondern auch die Wahrscheinlichkeit.

Wir haben also im Grunde genommen drei oder man könnte sogar fast sagen vier Kategorien, und eine davon ist die höchste, nämlich der Beleg. Hier haben wir uns auch schon in der Vergangenheit an dem allgemeinen Wissenschaftsverständnis orientiert, was meines Erachtens, aber vielleicht kann mich jemand eines Besseren belehren, unbestritten ist, dass nämlich die Replikation für den Beweis erforderlich ist. Die einmalige Beobachtung reicht in der Regel nicht aus, um etwas als bewiesen anzusehen, sondern das muss eben auch repliziert werden. Das führt dann zwangsläufig dazu, dass man eben nicht nur eine Studie hat, die einen bestimmten Effekt darstellt, sondern eben zwei. Das sahen in der Vergangenheit und sehen meines Wissens die Zulassungsbehörden immer noch ähnlich, indem sie eben auch von diesen regelhaften Situationen ausgehen. Man kann vielleicht sagen, dass in den letzten Jahren, was die Zulassung angeht, aus bestimmten Gründen vielleicht da etwas lockerer herangegangen wird - keine Ahnung, ob das stimmt -, aber es mehren sich Situationen, wo wir eben nur eine pivotale Studie haben. Das ist sicher so.

Jetzt haben wir in unserem Methodenpapier - auch schon in der Vergangenheit - geschrieben, dass der Regelfall für einen Beweis bedeutet, man braucht zwei Studien, aber es gibt eben auch Ausnahmen von dieser Regel, und haben in dem Zusammenhang auf eine Guideline referenziert, die EMA-Guideline. Das scheint aber den meisten Stellungnehmenden - fast alle Stellungnahmen haben das adressiert; deswegen kann ich jetzt auch gar keinen einzelnen oder einzelne Stellungnehmende ansprechen, sondern gebe das allgemein in die Runde - nicht auszureichen, sondern man wünscht sich eine stärkere Operationalisierung. Wann also genau reicht eine einzelne Studie aus, um zu sagen: „Das ist der Beweis“?

Ich würde gerne in diesem Zusammenhang, damit die Diskussion nicht völlig ausufert, eines zu bedenken geben - davon werden wir auch nicht abweichen können -: Es gibt sozusagen nicht ein Recht auf einen Beweis. - Wenn die Evidenz nur unsicher sein kann, aus welchen Gründen auch immer, weil es zum Beispiel eine extrem seltene Erkrankung ist - nur 20

Menschen auf der Welt haben diese Erkrankung -, dann wird man sich damit abfinden können, müssen, dass es eben nicht definitive Beweise in der Situation gibt. So ist es nun mal. Die Beweiskraft kann also nicht von den Ansprüchen abhängig gemacht werden, sondern sie ist nun mal so, wie sie ist, und obliegt einem allgemeinen Wissenschaftsverständnis.

Als Gedankengang: Wir würden Sie alle hier bitten und auffordern, vielleicht nicht alleine im Rahmen dieser Erörterung - das werden wir nicht schaffen -, Vorschläge zu machen, wie man das gut operationalisieren kann, zu sagen: Ja, hier eine Studie, die die und die und die und die Kriterien erfüllt, und zwar nicht die Rahmenbedingungen, nicht zu sagen, weil es eine seltene Erkrankung ist, reicht eine Studie - das ist Unsinn; das macht keinen Sinn; bitte das nicht; das brauchen wir nicht weiter zu diskutieren -, sondern eine Studie, die bestimmte Qualitätsmerkmale aufweist, weil sie zum Beispiel in der Lage ist, durchweg Konsistenzen nachzuweisen - dazu muss sie natürlich eine bestimmte Größe haben; das ist trivial -, oder eine Studie, die ganz tolle Effekte über alle Endpunkte konsistent hin zeigt, oder eine Studie, die multinational, multikontinental in Deutschland, in allen Bundesländern, in allen Landkreisen durchgeführt worden ist, die uns erlaubt, bestimmte Aussagen zu treffen. Wenn Sie uns Vorschläge für solche Kriterien machen, dann wären wir sehr dankbar. Aber bitte nicht den Anspruch formulieren! Den gibt es nicht. Den kann es auch nicht geben. Das macht auch gar keinen Sinn.

Moderator Jürgen Windeler: Soviel zur Einleitung. Soviel zu unserem Wunsch. Soviel zur Umsetzung Ihres Wunsches nach Operationalisierung. Jetzt sind wir gespannt.

Carsten Schwenke: Erst mal der Punkt zu zwei pivotalen Studien in der Zulassung: Es ist eben in der Zulassung nicht so, dass eine Studie repliziert werden soll, sondern dass es zwei pivotale Studien geben soll, also nicht zwei identische Studien auf dem gleichen Protokoll. Das ist, glaube ich, der Hauptunterschied zum Verfahren in der Nutzenbewertung, wo Sie ja gerade zumindest theoretisch zwei exakte Studien idealerweise haben würden oder wollen würden.

Stefan Lange: Erstes Missverständnis. Natürlich, Sie haben völlig recht: Es müssen nicht zwei Studien mit exakt dem gleichen Protokoll sein. Wollen wir auch nicht. Das ist zum Beispiel etwas, worüber man sich unterhalten kann: Was können sinnvolle Abweichungen sein?

Carsten Schwenke: Was ist ähnlich? Genau.

Bei der GMDS-Tagung hatte die Frau Sturz schon einmal ein paar Kriterien angesagt, auf deren Basis man auf einer pivotalen Studie einen Beleg bekommen könnte. Ich erinnere mich an so etwas wie ein sehr kleiner p-Wert. Die anderen Kriterien weiß ich nicht mehr. Sind das Kriterien, die das IQWiG als solches für akzeptabel hält?

Ralf Bender: Klar, sonst hätte das Sibylle da nicht so vorgetragen. Das sind Sachen, worüber wir jetzt nachdenken und wo wir ganz froh wären, wenn andere Leute sagen würden: Das ist so vernünftig, aber da fehlt vielleicht noch etwas - dann würden wir das ergänzen -, oder das ist so ausreichend. - Das war ja nicht eine ganz genaue Operationalisierung, sondern nur eine grobe Richtung, in die wir nachdenken. Reicht das, dass man das ganz grobe Kriterien nennt?

Reicht es zu sagen, kleiner p-Wert, oder muss man eine genaue Grenze angeben? Wie klein muss der p-Wert denn sein?

Carsten Schwenke: Kurze Rückfrage: Das heißt, ihr arbeitet daran, und wir bringen uns ein?

Ralf Bender: Ja.

Moderator Jürgen Windeler: Wir machen ja die Erörterung deshalb, weil viele Stellungnahmen gesagt haben, das, was wir an Operationalisierung beschrieben haben, reicht noch nicht aus, sondern es muss viel intensiver sein. Jetzt stellen wir uns vor, dass er, wenn er so eine Stellungnahme formuliert, eine Idee im Kopf hat, in welcher Beziehung es konkreter sein muss. Jetzt ist hier die Möglichkeit, zu sagen, das und das fehlt, das und das muss besser werden, das und das sind Kriterien, die unverzichtbar sind, das und das sind Kriterien, die völlig verzichtbar sind usw. Insofern machen wir das hier, um uns mit Ihnen auszutauschen, und zwar ein Stück weit jetzt.

Bernhard Wörmann: Ich bin dankbar, dass Sie das Thema aufgreifen. Aus der Onkologiesicht haben wir, glaube ich, bei den letzten 18 Nutzenbewertungen in 14 Fällen nur eine Studie gehabt. In vier Fällen gab es mehr als eine Studie. Insofern ist die Realität erst einmal so, dass wir mit einer Studie umgehen müssen. Das betrifft nicht Sie alleine, das betrifft uns auch in der Zulassung, aber vor allem bei Leitlinien, wo wir mit einer Studie umgehen müssen. In diesem Fall ist es vielleicht so, dass man sich der Realität anpassen muss, die man sich nicht immer so aussuchen kann.

Als zweites kritisches Kriterium würden wir die Qualität des Vergleichsarms als konstruktiven Vorschlag nehmen. Unser Beispiel ist Imatinib bei chronisch myeloischer Leukämie. Das ist der Türöffner für die gezielte Therapie gewesen. Da ist nur eine Studie mit über 1.000 Patienten gemacht worden, was relativ viel ist bei einer relativ seltenen Entität. Und dann kam ein dramatisch positiver Unterschied für Imatinib heraus, aber auch, weil der Vergleichsarm mit Interferon und ara-c sehr hoch dosiert war und es ein gegenüber anderen vorlaufenden Studien besseres Ergebnis im Vergleichsarm gab als erwartet. Der wichtige Punkt dabei ist: Ein p-Wert würde mir nicht reichen, weil es das Risiko öffnet, dass der Vergleichsarm untertherapiert wird, um einen größeren Unterschied herauszukriegen. Das heißt, das ist das wichtigste Kriterium neben der Effektstärke - das kann ich gut nachvollziehen -, auch dass man sich darauf einlässt, indirekte Vergleiche des Vergleichsarms zuzulassen, um nicht in die Falle hineinzulaufen, dass ein dramatischer Effekt aufgrund einer Unterdosierung des Vergleichsarms herauskommt. Das wäre ein zweites wichtiges Kriterium für uns.

Hans Wille: Ich weiß gar nicht, ob ich es etwas sagen darf, weil ich nämlich keine Vorschläge zur Verbesserung habe. Ich finde es schon ein Thema, das mich bewegt hat. In meinen Augen bedeutet Beleg mehr als Hinweis. Ich weiß nicht, welche Rolle das bei der Preisverhandlung spielt. Aber diese points to consider von 2001 sind ja eigentlich mit vielen unbestimmten Begriffen behaftet. Wenn ich an die Ticagrelor-Studie und an die Apixaban-Studie denke, es gab da zum Beispiel Ländereffekte, zum Beispiel USA bei Ticagrelor gegensätzliche oder völlig gegengerichtete, nicht ganz signifikante Effekte. Ich stelle mir vor, wenn das genau umgekehrt gewesen wäre, wenn das für die Population in Deutschland

nachgewiesen wäre. Ich kann eigentlich nur diese Fragen stellen. Ich habe natürlich keine Antworten darauf. Ich bin, glaube ich, auch nicht derjenige, der die beantworten kann.

Stefan Lange: Nur um Sie jetzt nicht bloßstellen zu wollen: Auch die AkdÄ gehörte zu denjenigen, die in ihrer Stellungnahme von uns gefordert haben, wir mögen das besser operationalisieren. Insofern wären wir für Vorschläge vielleicht auch im Nachgang, so es denn zeitgerecht ist, sehr offen.

Ich habe noch vergessen zu erwähnen, dass tatsächlich auch eine Organisation, die AWMF, unser bisheriges Vorgehen ausdrücklich begrüßt hat.

Monika Nothacker: Wobei wir auch gesagt haben, dass Sie die besonderen Anforderungen konkretisieren sollten.

Stephanie Rosenfeld: Ich möchte einen Vorschlag machen, den wir schon mehrmals gemacht haben auch in Stellungnahmen, und zwar dass man sich ansieht, wie viele Patienten die Studie beinhaltet und das in Bezug auf die Inzidenz in dem jeweiligen Land. Weil da haben wir auch gesehen, dass zum Beispiel Indikationen im Herz-Kreislauf-Bereich vielleicht zwei Studien hatten, die aber insgesamt einen weitaus geringeren Prozentsatz an den erkrankten Patienten ausmachen als vielleicht in seltenen Erkrankungen, wo wir dann eben auch Studien mit über 1.000 Patienten haben, wo wir halt nur eine Studie haben.

Einen zweiten Punkt möchte ich noch ansprechen. Sie sagen, Herr Lange, Sie möchten jetzt nicht grundsätzliche Fragen diskutieren, wo der Beweis und der Hinweis anfangen. Sie beziehen sich ja selbst in Ihrem Methodenpapier auf das EMA-Papier, das sich „Points to consider and application with one pivotal study“ nennt. Da ist ganz klar definiert, dass unter bestimmten Voraussetzungen eine Studie ausreichend ist. Da wird auch eine Nutzen/Risiko-Abwägung vorgenommen. Das ist natürlich europäisches Zulassungsrecht. Da ist jetzt die Frage, was höher steht und woran sich der Hersteller orientieren muss.

Stefan Lange: Es gibt ein Missverständnis. Ich habe nicht gesagt, dass wir nicht grundsätzliche Fragen besprechen können. Im Gegenteil! Wir wünschen uns das. Ich habe nur gesagt, man kann nicht sozusagen ein Recht reklamieren, nur weil es möglich sein muss, in ganz schrecklichen Situationen einen Beweis aussprechen zu können. Nein, ein solches Recht existiert nicht. Das macht auch keinen Sinn. Wir sind ein wissenschaftliches Institut und kein moralisches Institut. Das habe ich gesagt.

Das Zweite ist: Ja, wir haben diese Referenz gegeben. Aber dieses Papier - das hat auch Herr Wille gesagt - strotzt vor unbestimmten Begriffen. Und Sie wünschen sich von uns die Operationalisierung. Machen Sie uns doch mal Vorschläge! Reicht eine Studie mit 10.000 Patienten oder 5.000 oder 200 oder 20? Machen Sie uns Vorschläge, was „Konsistenz“ heißt. Das steht nämlich da alles drin. Da steht „Konsistenz“. Was ist „Konsistenz“? Sagen Sie es mir.

Susanne Teupen: Was aus unserer Sicht noch sinnvoll wäre, wäre der Endpunkt Lebensqualität, auch hier natürlich auf diesen Endpunkt zu powern. Natürlich könnte es ja auch ein Instrument sein, zu sagen, man hat ein krankheitsspezifisches und generisches Lebensqualitätsinstrument gewählt, die beide natürlich in die gleiche Richtung zeigen. Das

wäre vielleicht auch noch ein Anhaltspunkt, zu sagen, ob es operationalisierbar wäre, für eine Studie ausreichen könnte.

Christian Sieder: Wir haben hingewiesen auf die Wichtigkeit der Reproduktion für Studien. Es ist auch schon gesagt worden, dass die Zulassungsbehörden, insbesondere die ICH-Guideline, in der Reproduktion doch sehr Wert auf die Variabilität legen. Ich habe noch einen konstruktiven Vorschlag, wie man auch mit weniger Evidenz zu einem Beleg kommen kann, nämlich indem man einfach die Evidenz, die in benachbarten Gegenden vorliegt, einbezieht, sprich zum Beispiel die Phase-II-Studien, die in der Nutzenbewertung überhaupt nicht vorkommen, oder die Konsistenz zu Studien, die zum Beispiel einen anderen Komparator verwenden als die zweckmäßige Vergleichstherapie. Teilweise ist es auch so, dass in der Nutzenbewertung noch nicht einmal die gesamte Studienpopulation berücksichtigt wird, sondern nur die Subpopulation, die der Zulassung entspricht. Auch da wäre es durchaus eine Möglichkeit, zu sagen: Ich kann schon aus benachbarten Populationen, die in der Studie sind, eine gewisse Evidenz herausziehen und muss die nicht einfach von vornherein ausblenden.

Thomas Kaiser: Ich habe mehrere Punkte und fange mit dem letzten an. Es ist erst einmal grundsätzlich falsch, dass wir Phase-II-Studien nicht berücksichtigen. Wenn wir die Frage beantworten, berücksichtigen wir die. Es gibt mehrere Beispiele, wo die schon berücksichtigt worden sind. Was ich allerdings mit einem Vergleich A versus C für einen Vergleich A versus B anfangen soll, ist mir schleierhaft, es sei denn, ich mache es über den indirekten Vergleich. Ich kann das ja nicht beliebig machen.

Dass wir Zulassungspopulationen berücksichtigen, also dass die Zulassung, nachdem sie durchgeführt wurde, gegebenenfalls eine Einschränkung vornimmt, ist, glaube ich, für die Bewertung nur sachgerecht. Wenn man mal auf Ticagrelor schaut - das Beispiel, das Herr Wille eben angeführt hat -, wenn wir da die Gesamtstudie betrachten würden inklusive der schlechteren Ergebnisse in Nordamerika, dann könnten Sie hier nicht mehr von einem Beleg für die deutsche Situation sprechen. Für die deutsche Situation sprechen wir aber von einem Beleg, weil eben diese Kombination mit einer hohen Aspirindosis in Deutschland keine Rolle spielt und die Zulassung auch so ausgesprochen ist. Das heißt, es ist völlig sachgerecht, sich für Deutschland das genau anzuschauen.

Sie haben gerade gesagt, hier gibt es ein höheres Zulassungsrecht. Wir stellen ja nicht die Zulassung infrage. Hier geht es um etwas ganz anderes: Welche Sicherheit haben Sie in der Nutzenbewertung? Das ist doch erst einmal völlig unabhängig davon, welche Sicherheit Sie in der Zulassung haben. Die Zulassung und auch diese Guideline machen es natürlich für den gesamten europäischen Kontext. Wenn Sie sich in Europa die verschiedenen Arten von Behandlungen anschauen, dann kann doch die Aussage für Deutschland nicht zwangsläufig die gleiche Sicherheit wie die Aussage für insgesamt Europa haben. Die Diabetesbehandlung in Europa ist heterogen. Ob Schulungen durchgeführt werden oder nicht, ist in vielen Ländern mehr oder weniger Standard. Insofern müssen Sie doch für die deutsche Situation betrachten, welche Antworten die Daten für die deutsche Situation geben, und Sie müssen nicht betrachten, welche Antworten die Daten für die amerikanische Situation geben, und auch nicht, welche Antworten sie für die britische Situation geben. Sie haben ja auch eine Vergleichstherapie, die Sie für Deutschland anwenden sollen. Das heißt, das reflexhafte Reflektieren immer wieder auf die Zulassungssituation, auf die europäischen Guidelines kann für bestimmte Dinge hilfreich sein, aber man muss sie doch selbstverständlich für die

Nutzenbewertung anwenden. Das mit Inhalt zu füllen, das ist unsere Aufforderung an dieser Stelle, also nicht wieder zu sagen, nehmt doch einfach das, was in Europa gesagt wird, sondern selber einen Vorschlag zu machen, was das für die Nutzenbewertung bedeutet. Darauf haben wir bislang kaum Antworten bekommen.

Friedhelm Leverkus: Ich denke schon, dass man die Zulassungssituation durchaus als Ausgangspunkt nehmen kann. Es gibt natürlich Abweichungen - da haben Sie vollkommen recht -, wenn man in der Zulassungssituation nicht die zweckmäßige Vergleichstherapie hat oder man hat im Prinzip andere Endpunkte drin. Aber ist es halt gegeben, stellt die Zulassung sich natürlich auch die Frage: Ist die Datenlage so, dass ich das Medikament den Menschen zutrauen kann? Ist es so, dass ich einen Beleg dafür habe, dass das Medikament wirksam ist? Häufig ist es so, dass Sie eine ähnliche Frage haben. Sie sind ja im Prinzip auch damit gestartet. Sie haben gesagt, ich brauche zwei Studien, wie die Zulassungssituation halt war. Ich kann mich noch an die Diskussion 2008 erinnern, das erste „IQWiG im Dialog“. Das ist meine Startposition, von der man sicherlich mal ausgehen und schauen sollte, wir können es halt nicht gebrauchen, weil wir jetzt einen anderen Komparator haben oder weil, wie Sie gesagt haben, Herr Kaiser, die Zulassungssituation in Deutschland eine andere ist wie in Studien beschrieben. Aber man sollte das meines Erachtens auch mit berücksichtigen.

Thomas Kaiser: Ich halte das, was Sie gerade gesagt haben, nicht für richtig. Die Zulassungsbehörden stellen nicht fest, dass es einen Beweis gibt, dass das zugelassene Medikament besser ist als die zweckmäßige Vergleichstherapie, selbst wenn die Zulassungsstudie im Vergleich zu dieser Therapie durchgeführt worden ist. Was die Zulassungsbehörde feststellt, ist, ob dieses konkrete Präparat mehr Nutzen als Schaden hat - Nutzen größer Risiko, das ist ja die Feststellung -, aber nicht, ob es einen Beweis, also einen Beleg dafür gibt, dass es besser ist als die Vergleichstherapie. Ich kenne keinen einzigen EPA, wo das steht. Deswegen, glaube ich, sollte man auch nicht denken, dass, wenn die Zulassung allein auf Basis einer solchen Studie oder der pivotalen Studie - es sind ja noch mehrere Daten, die da eingereicht werden - durchgeführt wird, das gleichbedeutend ist mit einem Beleg bezüglich eines Zusatznutzens gegenüber der zweckmäßigen Vergleichstherapie. Das ist schlicht und einfach keine Fragestellung, die die Zulassung beantwortet.

Ch.-Markos Dintsios: Eingehend auf die Evidenz, die hier vorliegt: Bereits in der Verordnung steht wortwörtlich, dass die frühe Nutzenbewertung - mit der befassen wir uns als Arzneimittelhersteller nun mal - auf Basis der Zulassungsstudien zu erfolgen hat, unabhängig davon, dass die Fragestellung eine andere ist. Ich gebe Herrn Kaiser vollkommen Recht. Die Fragestellung in der Zulassung ist eine vollkommen andere. Nichtsdestotrotz ist aber die vorliegende Evidenz genau diejenige. Und weil man mit dieser Evidenz arbeiten muss, weil man die Kriterien der Zulassungsbehörden erfüllen muss, ist das eine gesetzte Rahmenbedingung. Es ist ein Datum. Das können wir leider nicht ändern, ob wir wollen oder nicht. Deswegen würde ich die Qualität der Evidenz zwar zur Disposition stellen für die entsprechenden Fragestellungen, aber ich würde sie nicht negieren, denn es gibt nun mal eine Grundmaxime in der evidenzbasierten Medizin: Bewerte die vorhandene Evidenz. - Also kann ich nicht sagen, weil die Evidenz für eine andere Fragestellung generiert wurde als die, die ich hier gerne beantwortet haben will, ignoriere ich sie vollends.

Zum Zweiten: Stefan Lange hat es, glaube ich, damals beim „IQWiG im Dialog“ gesagt, das Problem mit den Zulassungspopulationen. Ich nehme mal die Causa Fingolimod, weil Herr

Sieder - er ist gerade draußen - sich vorhin gemeldet hat. Da haben wir das Problem, die Zulassungspopulation war enorm klein am Ende, aber wir haben auch das umgekehrte Problem, die Zulassung lässt auch Analogieschlüsse zu. Sie erweitert manchmal sogar die intendierte Population. Das darf man auch nicht unterschlagen. Also haben wir leider ein Dilemma, aus dem wir nicht so leicht herauskommen unter den gegebenen Rahmenbedingungen, die wir und übrigens auch das IQWiG nicht gesetzt haben, sondern die uns von der Politik vorgegeben wurden. Man kann das sehen, wie man will. Ich erachte das nicht als einen glücklichen Zustand - ich mache daraus keinen Hehl -, aber es ist nun einmal so.

Ein kurzer Kommentar zu Frau Teupen: Ich würde die Wertigkeit einer einzelnen Studie nie alleine an der Endpunktqualität festmachen, ob jetzt die Lebensqualität gepowert wird oder nicht. Die Lebensqualität ist ein Konstrukt. Das ist ein sozialempirischer Konstruktivismus, in dem wir uns bewegen. Die ist höchst patientenrelevant, aber an dieser einzelnen isolierten Endpunktqualität die Wertigkeit einer Einzelstudie für einen Beleg festzumachen, wäre mir persönlich extrem kurzgegriffen.

Stefan Schinzel: Ich habe zwei Kommentare, einen nicht ganz ernstgemeinten: Das Risiko, dass zwei Studien falschpositiv in dieselbe Richtung sind, ist 0,025 zum Quadrat. Das ist sicher eine Zahl, über die man in dem Zusammenhang nachdenken muss, aber es ist sicherlich nicht das alleinige Kriterium.

Der zweite Punkt, den ich sehr viel ernster meine und der hier schon angeklungen ist, ist: Man ist häufig in der Situation, dass man nur eine pivotale Studie hat, aber vielleicht zusätzlich indirekte Vergleiche konstruieren kann, sei es aus placebokontrollierten Studien, sei es aus Phase-II-Studien. Herr Dr. Kaiser hat ja schon angedeutet, dass man da unter Umständen den falschen Komparator hat. Für uns stellt sich halt die Frage, wenn wir uns Tabelle 2 aus Ihrem Vorschlag anschauen, was es für einen zusätzlichen indirekten Vergleich gibt, der eine einzelne pivotale Studie unterstützt.

Stefan Lange: Darauf kann ich direkt antworten. Ich wollte gerade sagen, einen halben Punkt.

Erstens fand ich Ihren ersten Vorschlag auch ernsthaft. Darüber muss man sicher nachdenken. Wenn man zum Beispiel ein Kriterium p-Wert formuliert, dann wird man sich daran orientieren. Was bedeuten denn zwei falschpositive Studien? Sie sind ja schon sehr konservativ, indem Sie schon das Alphahalbe zweiseitig nehmen. Aber das ist okay. Das ist ja Ihr Vorschlag.

Ganz wichtig ist - ich glaube, das kann man nicht oft genug sagen, damit das nicht falsch stehen bleibt oder irgendjemand das im Protokoll herauspickt, was Herr Leverkus gerade gesagt hat -: Nein, Zulassung und frühe Nutzenbewertung sind völlig unterschiedliche Fragestellungen. Darüber brauchen wir, glaube ich, nicht weiter zu reden. Das ist ganz wichtig.

Ich würde nur ganz gerne als Möglichkeit aufgreifen und wieder betonen, was wir an anderer Stelle auch schon gesagt haben: Solange wir keine allgemeingültigen Regeln formuliert haben, ist es immer an Ihnen, im Dossier auch andere Evidenz - die Möglichkeit besteht ja;

das ist ausdrücklich vorgegeben auch durch die Arzneimittelnutzenverordnung - heranzuziehen, zum Beispiel einen indirekten Vergleich, um eben Ihre Aussagen zu stützen. Wir kommen ja unter TOP 5 noch zu der Evidenzaufwertung durch indirekte Vergleiche. Ja, ausgeschlossen ist das nicht, auch für diese Frage. So weit würde ich mich jedenfalls aus dem Fenster lehnen wollen. Es ist immer die Frage: Wie überzeugend ist es im Dossier nachgewiesen? Nur zu sagen: „Ich habe jetzt hier einen anderen Komparator. Da habe ich ein ganz schönes Ergebnis. Der hat zwar nichts mit der zweckmäßigen Vergleichstherapie zu tun, aber trotzdem schön“, das ist sicher nicht ausreichend, sondern da wird man schon methodisch ein bisschen geschickter den Bezug zur tatsächlichen zweckmäßigen Vergleichstherapie herstellen müssen. Aber das ist, denke ich, eine wunderbare Möglichkeit.

Also haben wir jetzt schon vier gute Kriterien aufsammeln können. Das war einerseits die Frage der Ausgestaltung des Komparators in der konkreten Studie. Das war die Endpunktqualität, die, glaube ich, auch nicht ganz unwichtig ist. Das war die Frage des p-Wertes Größenordnung, zum Beispiel die Situation, dafür zwei falschpositive Studien heranzuziehen. Oder eben die Möglichkeit des indirekten Vergleichs.

Stephanie Rosenfeld: Ich habe auch noch einen Vorschlag, wie man die Frage lösen könnte, ab wann eine einzelne Studie ausreichend ist, und zwar könnte man das in Zusammenarbeit mit Ethikkommissionen entscheiden, weil es wird immer wieder gerade bei seltenen Erkrankungen Indikationen oder Situationen geben, in denen es nicht ethisch ist, eine zweite ähnlich gelagerte Studie zu machen. Es mag sein, dass es viele Fälle gibt, wo man tatsächlich mehr als eine Studie machen könnte. Es gibt auch Fälle, wo es tatsächlich unethisch wäre, eine zweite Studie in diesem Setting zu machen. Da könnte man zum Beispiel Ethikkommissionen heranziehen, man könnte auch das BfArM heranziehen, um diese Werteentscheidung zu treffen, ob es tatsächlich ethisch ist, Patienten eine Therapie vorzuenthalten, die nachweislich einen Vorteil gezeigt hat.

Moderator Jürgen Windeler: Ich muss noch einmal deutlich machen, dass dieses Argument - Stefan Lange hat es auch bereits gesagt - meiner Ansicht nach völlig an der Sache vorbeigeht. Es geht überhaupt nicht um die Frage, ob es ethisch ist, eine zweite Studie zu machen. Es mag sehr wohl sein, dass es unethisch ist, eine zweite Studie zu machen. Das ist doch völlig in Ordnung. Dem würden wir uns auch in keiner Weise… Wir fordern nicht zwei Studien. Es ist aber auch klar, dass mit einer Studie die Unsicherheit eines Ergebnisses höher ist als bei zwei, drei und vier Studien. Das einzige, was wir sagen, ist: Wenn nur eine Studie da ist, auch dann, wenn man nur eine Studie machen kann, aus welchen Gründen auch immer, dann ist das Ergebnis unsicherer, als wenn es mehrere sind. Das ist das einzige, was wir sagen. Wir fordern nicht, um das deutlich zu machen, zwei, drei Studien und Replikationen in unethischen Situationen - das ist an den Haaren herbeigezogen, falls Sie das gemeint haben sollten -, sondern was wir sagen, ist: Für eine Studie gibt es in der Regel keinen Beleg. Das ist auch bei seltenen Erkrankungen so, weil einfach die Studie Unsicherheit deutlich macht und dann in einem Hinweis oder ähnlichen Dingen abgebildet ist.

Ilona Krug: Meine Frage zielte auch genau darauf. Ich will noch einmal das Beispiel Vemurafenib nennen. Das ist genau das, was Sie gerade sagten. Die wurde aus ethischen Gründen praktisch nach dem zweiten Datenschnitt crossoverdesigned, weil man den Patienten nicht mehr zumuten konnte, praktisch im Vergleichsarm zu verbleiben. Wie sehen Sie das?

Da hat man ja als Hersteller niemals die Chance, einen Beleg zu bekommen? Da sagen Sie, das ist dann halt so?

Moderator Jürgen Windeler: Das ist dann halt so, es sei denn, wir erarbeiten gemeinsam - wir waren ja schon auf dem Weg - Kriterien, die möglicherweise auch solche Situationen mit abgreifen könnten. Die Kriterien müssten dann sehr überzeugend und allgemein anwendbar sein. Ich persönlich bin ein bisschen skeptisch, ob es in dieser konkreten Situation die geben kann. Aber wenn man solche Situationen allgemein erhebt, die zum Beispiel auch in der Konsistenz von Ergebnissen über verschiedene Anwendungssituationen, sprich verschiedene Zentren oder Ähnlichem, liegen könnten, kann man so etwas machen. Aber im Grundsatz: Wenn man es nicht hinbekommt - das ist doch der einzige Punkt, über den wir reden -, über welche Wege auch immer - das mögen externe ernstzunehmende Hinderungsgründe sein, die wir alle nicht ändern können -, eine gewisse Ergebnissicherheit zu erreichen, wird es keinen Beleg geben, kann es gar nicht.

Friedhelm Leverkus: Ich glaube, ich bin eben falsch verstanden worden. Ich will nicht sagen, dass es immer die gleiche Fragestellung ist, aber in meinem einfachen biometrischen Denken kann ich mir Situationen vorstellen, wo ich die gleiche Studie, die gleiche Aussage habe. Ich habe den ZVT, habe im Prinzip den Komparator, mache eine Überlegenheitsstudie. Und diese eine Studie wird von der Zulassungsbehörde als Beleg anerkannt. Da wird eine Entscheidung getroffen. Und wenn ich die Studie beim IQWiG einreiche - das ist genau das Gleiche -, wird es halt nicht anerkannt. In dieser Situation wäre es, finde ich jedenfalls, auch okay, wenn man mal guckt, was die Zulassungsbehörden dazu sagen. Darüber hinaus gebe ich Ihnen natürlich Recht, dass man über weitere Kriterien nachdenken sollte.

Stefan Lange: Dann haben wir Sie nicht falsch verstanden, Herr Leverkus. Das ist trotzdem falsch. Auch wenn es nur diese eine Studie gäbe, die es für die Zulassung ja gar nicht gibt - das haben wir ja schon festgestellt; die Zulassung zieht auch andere Daten heran, die wir teilweise auch heranziehen, aber teilweise eben auch nicht -, ist es trotzdem nicht die gleiche Fragestellung. Auch die Konsequenz ist nicht die Gleiche. Natürlich haben Sie recht, dass die Zulassung zu der Marktzulassung Ja oder Nein sagt. Wir sagen nicht Ja oder Nein, wir sind gefordert, das Ausmaß und die Wahrscheinlichkeit des Zusatznutzens zu beschreiben. Das ist etwas völlig anderes, als ob man ein Medikament für den Marktzugang zulässt oder nicht. Das ist etwas völlig anderes. Das ist nicht das Gleiche. Die Konsequenz ist eben nicht eine binäre Entscheidung Ja oder Nein, sondern die Konsequenz ist dann Grundlage von Preisverhandlungen. Bitte hören Sie damit auf! Es hilft einfach nicht.

Bernhard Wörmann: Ich glaube, das Beispiel von Vemurafenib bildet sehr gut ab, was wir inzwischen zusammen diskutiert haben. Vemurafenib ist, glaube ich, nahe an einem Durchbruchsmedikament für das Melanom gewesen. Ich glaube, es ist deswegen gut, weil es eine sehr hohe Effektstärke hatte, weil Überlebenszeit verbessert wurde, Remission verbessert wurde. Insofern ist das, glaube ich, nahe dran an dem, was wir diskutieren. Und der Vergleichsarm, den man gegen Chemotherapie verglichen hat, war zu der Zeit sehr korrekt. So würden auch die anderen Kriterien passen. Vemurafenib ist, glaube ich, aber auch ein Beispiel, dass man es in der Gesamtheit sehen muss, weil das Risiko für die Zweitkarzinome in der ersten Bewertung mit fast 25 % so hoch gewesen ist. Wäre das ein Minus gewesen, glaube ich, wäre wahrscheinlich auch mit einer Studie kein „erheblich“ dabei herausgekommen. Das ist, glaube ich, schon ziemlich nahe dran.

Ich möchte Sie ausdrücklich noch einmal unterstützen. Ich glaube nicht, dass es einen Minderheitenschutz von seltenen Erkrankungen geben kann. Wir lernen, dass, wenn gut multinational organisiert, auch relativ seltene Erkrankungen sich heute ganz gut in Phase-III-Studien abbilden lassen, zumindest in einer hochkarätigen Studie.

Ilona Krug: Aber dann eine Frage. Vemurafenib wurde ja dann abgebrochen. Wie wollen Sie denn damit umgehen? Deswegen meine Frage an das IQWiG zurück: Habe ich es richtig verstanden, dass im Grunde für solche Sachen, also Sonderfälle, wo nur eine Studie oder eine Studie aus ethischen Gründen abgebrochen wird, Vorschläge erarbeitet werden können, wie damit umgegangen wird?

Ch.-Markos Dintsios: Ich wollte jetzt nicht Herrn Lange das Wort nehmen.

Anschließend an das, was wir gehört haben: Vielleicht könnte man noch ein weiteres Kriterium einbauen, eine Art intentionales Kriterium, das sozusagen den Hinweis ein bisschen qualifizieren würde, und zwar - ich bin gleich fertig - ob in dieser Hinsicht etwas machbar ist oder nicht. Ich bin der Auffassung, auch bei seltenen Erkrankungen kann man mehrere Studie machen, aber es gibt manchmal Fälle - wir hatten ja eine Generkrankung, die bei Portugiesen auftauchte, im Rahmen auch von Orphan-Drug-Bewertung -, wo dann annähernd die Gesamtheit Deutschlands rekrutiert wurde. Da wird es ein bisschen schwierig. Man könnte zwei Studien machen. Die zwei Studien wären Minimalstudien, statistisch überhaupt nicht auswertbar. Dann würde ich durchaus vorschlagen, eine Feasibility-Frage zu stellen. Also ich würde den Hinweis noch unterteilen. Es gibt ja die Möglichkeit, einen Beleg über zwei Studien zu liefern. Aber es sind wirklich Konstellationen, die nichts mit Ethikkommissionen zu tun haben, die haben ganz einfach mit dem Zahlenwerk zu tun, wo eine zweite Studie ganz einfach extrem schwer, wenn nicht sogar nicht machbar wird.

Moderator Jürgen Windeler: Ich kann nicht an mich halten. Um es noch einmal zu sagen: Selbst wenn wir in einer Situation wären, wo für eine irgendwie geartete medizinische Intervention aus ethischen oder sonstigen Gründen keine Studie gemacht werden kann - mag es ja geben -, keine vergleichende Studie gemacht werden kann, selbst wenn wir dort wären, dann heißt das einfach nur, dass wir über diese Intervention ziemlich wenig wissen. Und dieses „ziemlich wenig wissen“ heißt, es kann kein Beleg werden. Da ist es völlig egal, ob man diese Studien machen könnte und machen kann und nicht jemand, den ich bezahlen soll, oder aus ethischen Gründen oder ob man die Patienten nicht findet oder ob sie so selten sind. Für diese Frage, ob das Ergebnis dieser Intervention einen gewissen Sicherheitsgrad oder einen gewissen Unsicherheitsgrad hat, spielt das alles keine Rolle. Die einzige Frage ist: Die Studie ist nicht da.

Stefan Lange: Ich glaube, das Vemurafenib-Beispiel kann in gewisser Weise paradigmatisch werden und könnte sich vielleicht eignen, zu versuchen, vernünftige Kriterien abzuleiten. Ob sie dann tatsächlich greifen würden und das der Fall ist, bin ich noch nicht so sicher.

Ich will nur eines sagen: Ich glaube, dass das falsch ist, was Sie gesagt haben: Die Studie ist nicht aus ethischen Gründen abgebrochen worden. Hier war wie in vielen onkologischen Indikationen im Protokoll angelegt, dass das Crossover ermöglicht wird. Und dann hat es einen Datenschnitt, zwei Datenschnitte, drei Datenschnitte gegeben. Das wäre mir jetzt ganz wichtig, auch wenn das mit dem Thema gar nicht so furchtbar viel zu tun hat, nämlich immer

das Bemühen irgendeiner Ethik: Dass wir Studien so konstruieren, wie sie gerade konstruiert werden, und dass man sie möglichst frühzeitig im Grunde genommen auf ihre Aussagekraft hin abschneidet, das kann man von seiner Ethik auch sehr stark hinterfragen. Das wird übrigens in der Literatur auch getan. Wir sind, glaube ich, dringend geboten, mal endlich über diese Ethik, dieses bewusste Nichtwissen oder die Generierung von Nichtwissen nachzudenken, Generierung von völlig unpräzisen Studienergebnissen, die hinterher die Anwender in ganz blöden Fahrwassern lassen. Wir sind uns ja einig, dass dieses große Probleme auch bei der Auswertung, bei der Interpretation der Ergebnisse liefert. Wir haben darüber diskutiert, was das zum Beispiel für die Interpretation von unerwünschten Ereignissen bedeutet, die im Augenblick meines Erachtens völlig verheerend, übrigens auch zum Nachteil der Präparate ist, was lustiger Weise bisher noch niemanden interessiert hat. Aber bitte, wir können ja so weitermachen wie bisher. Bitte hören Sie mit der Ethik auf!

Moderator Jürgen Windeler: Jedenfalls in diesem konkreten Kontext.

Ch.-Markos Dintsios: Ich glaube, Sie haben mich vorhin falsch verstanden. Ich habe nicht argumentiert in solchen Konstellationen pro Beleg. Ich habe klipp und klar gesagt, auch wenn ich nicht native sprechend bin, dass man dann die Kategorie „Hinweis“ noch einmal mit einem Kommentar belegen muss, nämlich war nicht anders machbar. Es bleibt immer noch ein Hinweis, Herr Windeler. Das habe ich gesagt. Ich möchte nicht, dass Sie mir meine Aussagen hier in ihren Inhalten drehen. Das gefällt mir nämlich überhaupt nicht. Das möchte ich gerne protokolliert haben.

Moderator Jürgen Windeler: Da wir ein Wortprotokoll machen, Herr Dintsios, ist auf jeden Fall die Protokollierung jedes einzelnen Wortes, jedes einzelnen Kommas sichergestellt.

Zum Zweiten: Selbstverständlich ist es so, wenn Sie die IQWiG-Berichte lesen, was Sie ja tun, dass wir dann wie üblicherweise im Detail beschreiben, wie wir zu unseren Wahrscheinlichkeitsfestlegungen kommen. Insofern ist die Situation hier: Wir haben keine Studie. Und wir haben einen Hinweis. Die Frage, dass es diese Studie aus irgendwelchen Gründen nicht gegeben hat, die wir auch gar nicht überprüfen können, wird unseren Lesern wahrscheinlich nicht so viel weiterhelfen.

Guido Skipka: Um konkret auf den inhaltlichen TOP zurückzukommen: Mir stellt sich gerade schon eine Frage. Ich finde den Punkt wirklich schwierig. Wenn wir jetzt einmal die Situation betrachten, man hat jetzt nicht extrem seltene Ereignisse und muss überlegen, ob man überhaupt auch nur eine Studie hinbekommen kann, und man ist in einer komfortableren Situation mit höheren möglichen Fallzahlen: Was ist denn eigentlich besser? Dass ich zwei kleine Studien mache, vielleicht mit zweimal 100 Patienten, also in der einen Studie 100 und in der anderen Studie 100 Patienten, oder sollte ich eine große Studie machen mit 1.000 Patienten, die natürlich wesentlich präziser ist, wo ich auch eine höhere Power habe, um Subgruppenanalysen durchzuführen? Da stellt sich für mich schon die Frage: Ab wann wird unser wichtiges Kriterium der Replikation schwächer? Ich weiß nicht, ob meine Frage verständlich war.

Moderator Jürgen Windeler: Gibt es dazu Wortmeldungen?

Carsten Schwenke: Noch eine spezielle Frage zum Bestandsmarkt. Da ist es ja häufig so, dass man gerade in der Onkologie eine klinische Studie hat und dann zum Beispiel Register-, Kohortenstudien, was auch immer man an Daten gesammelt hat. Bisher ist es, zumindest wenn ich es richtig verstanden habe, so, dass man die RCT heranzieht für den Zusatznutzen und die zusätzlichen Daten zwar interessant findet, aber nicht zwangsläufig mit in die Bewertung einbezieht. Wäre es denn eine Möglichkeit, wenn es denn gute Studien sind, also eine gute Kohortenstudie, ein wirklich exzellentes Register, diese zu nutzen, um die Evidenzgrundlage ein bisschen zu erhöhen?

Stefan Lange: Es gibt die Möglichkeit - die ist auch vorgesehen; ich weiß nicht, welcher Abschnitt das im Dossier ist; es hat irgendeine Nummer, 4.3.2 -, wo Sie zusätzliche Evidenz mit heranziehen können, um Ihre Aussagen zu stärken. Der Teufel steckt jedoch bekanntlich im Detail. Was ist eine gute Kohortenstudie? Sie haben jetzt auch den Begriff des Registers verwendet. Das eine hat, glaube ich, mit dem anderen nicht zwangsläufig etwas zu tun. Man kann unter Umständen vielleicht aus guten Registern - was ist ein gutes Register? - versuchen, gute Kohortenstudien zu generieren. Keine Ahnung. Ich glaube, das wird mitunter sehr schwierig, aber ausgeschlossen ist das nicht. Nur noch einmal: Es ist dann in der Verantwortung desjenigen, der das Dossier erstellt, sehr überzeugend darzulegen, warum genau diese Studie, die nicht Evidenzstufe 1 entspricht, trotzdem eine so starke Unterstützung für die vielleicht vorliegende Studie ist, damit man sagen kann: Das ist der endgültige Beweis. - Ich halte es nicht für ausgeschlossen, ich halte es aber zugegebenermaßen auch nicht für sehr wahrscheinlich.

Thomas Kaiser: Um ein bisschen Realismus mit einfließen zu lassen: Auch in diesen Abschnitten ist natürlich der Hersteller dazu aufgefordert, das systematisch zu machen, also sich nicht eine beliebige Kohortenstudie zu nehmen, sondern systematisch die vorhandene Evidenz zu sichten und darzustellen. Es kann also nicht die eine positive unterstützende - den Wunsch danach lese ich manchmal da heraus, gar nicht bei Ihnen, sondern generell - Evidenz für diesen einen Endpunkt sein - das gilt für Registerstudien wie für indirekte Vergleiche gleichermaßen -, sondern das muss selbstverständlich systematisch für die gesamte Fragestellung gemacht werden. Wenn ich an der einen Stelle durch diese Evidenz auch etwas unterstützen kann, kann das natürlich für einen anderen Endpunkt bedeuten, dass ich hier zusätzlich eine Evidenz bekomme, die ein anderes Endpunktergebnis infrage stellt. Das kann ich ja nicht beliebig einfach nur dafür machen, wofür ich das gerne hätte.

Moderator Jürgen Windeler: Dann schließe ich den Tagesordnungspunkt.

Ich würde jetzt gucken, ob wir mit Tagesordnungspunkt 4 in den nächsten 15 bis 20 Minuten durchkommen. Auf jeden Fall machen wir dann die Mittagspause. Wenn wir uns mit dem TOP 4 irgendwo verzetteln und verhaspeln oder es sehr kompliziert wird, dann müssen wir ihn unterbrechen und die Mittagspause einschieben.

Wir kommen zum

Tagesordnungspunkt 4:

Anwendung und Interpretation von Prädiktionsintervallen

Ralf Bender: Wir haben mit unserem Entwurf für das Methodenpapier 4.1 die Prädiktionsintervalle neu eingeführt. Dazu gab es mehrere Stellungnahmen. Ich glaube, die basieren zum Teil auf Missverständnissen. Das würde ich gerne hier diskutieren.

Vorab: Wir verwenden die Prädiktionsintervalle insbesondere als Tool, um in einer Situation, die zu heterogen ist, als dass man sinnvoll eine Meta-Analyse durchführen könnte, trotzdem auf qualitativer Art und Weise, unterstützt durch ein quantitatives Tool, zu Nutzenaussagen, Beleg, Hinweis, Anhaltspunkt zu kommen. Da spielt der Begriff der Gleichgerichtetheit eine Rolle, jetzt die Prädiktionsintervalle, um festzustellen, ob in einer heterogenen Situation, wo es nicht sinnvoll ist, einen gepoolten Effektschätzer zu haben, die aber doch dann bei deutlicher Gleichgerichtetheit trotzdem einen Beleg ergibt oder bei mäßiger Gleichgerichtetheit eben einen Hinweis ergibt. Da verwenden wir Prädiktionsintervalle.

Zwei Aspekte dazu: In der Stellungnahme der Firma Lundbeck wird geschrieben - ich lese es vor -:

> „Außerdem ist es notwendig, den bestehenden Widerspruch aufzulösen, Gleichgerichtetheit nur dann darzustellen, wenn keine Meta-Analyse durchgeführt werden kann, jedoch andererseits Prädiktionsintervalle in meta-analytischen Berechnungen zu ermitteln."

Ich bin nicht ganz sicher, ob ich dieses Argument richtig verstanden habe. Eine mögliche Interpretation wäre, dass hier nicht verstanden wird, wieso man in einer Situation, wo man keine Meta-Analyse durchführt, trotzdem Prädiktionsintervalle berechnet. Wenn das die Frage ist, antworte ich darauf wie folgt: Das Prädiktionsintervall ist eine grafische Darstellung der Heterogenität. Wir können natürlich sehr wohl, auch wenn es aufgrund der zu großen Heterogenität nicht sinnvoll ist, einen gepoolten Effektschätzer darzustellen, einen Forest Plot darstellen, eben ohne den gepoolten Effektschätzer. Und das Prädiktionsintervall ist dann eine grafische Darstellung der Heterogenität in diesem Forest Plot. Das ist natürlich dann sinnvoll, wenn es nicht sinnvoll ist, einen gepoolten Effektschätzer darzustellen. Frage ist, ob die Firma Lundbeck mit dieser Antwort zufrieden ist, ob das das erfüllt, oder ob in diesem Absatz, den ich vorgelesen habe, doch noch etwas anderes steckt, was ich vielleicht nicht verstanden habe.

Markus Kessel-Steffen: Sie haben die Frage richtig verstanden. Das heißt, in jedem Fall die Daten meta-analytisch zusammenzufassen, um dann den Forest Plot darzustellen, ungeachtet gegebenenfalls aus Ihrer Sicht zu großer Heterogenität mit der Konsequenz, dass dann kein Gesamteffektschätzer dargestellt werden kann.

Ralf Bender: Ich weiß nicht, ob ich das richtig verstanden habe. Natürlich verwenden wir, um die Prädiktionsintervalle zu berechnen, meta-analytische Formeln, genauso wie man auch das I^2 berechnet. Wir führen quasi formal schon eine Meta-Analyse durch, stellen nur den gepoolten Effektschätzer nicht dar, weil die Heterogenität zu groß ist.

Markus Kessel-Steffen: Genau.

Ralf Bender: Das halten wir so für sinnvoll. Ist die Frage beantwortet?

Markus Kessel-Steffen: Ja.

Ralf Bender: Dann schaffen wir sogar noch den zweiten Punkt. Hier scheint ein Missverständnis vorzuliegen, zum Beispiel in der Stellungnahme von Pfizer, die hier schreiben:

> „Es ist nicht begründet, warum das IQWiG von der Festlegung der Beleglage auf Basis von gleichgerichteten Effekten aus dem Methodenpapier 4.0 abgewichen ist und diese Verschärfung eingeführt wurde."

Hier werden also, glaube ich, die Prädiktionsintervalle als Verschärfung verstanden. Das sehen wir nicht so. Die Prädiktionsintervalle stellen keine Verschärfung dar, sondern es ist eine genauere Operationalisierung. Vorher haben wir im Methodenpapier 4.0 grobe Kriterien dargelegt, wann wir von Gleichgerichtetheit sprechen. Jetzt, da wir ein neues quantitatives Tool zur Verfügung haben, sind wir in der Lage, sogar in einer besseren Operationalisierung darzulegen, wann etwas deutlich oder mäßig gleichgerichtet ist, und das eben so, dass Subjektivität soweit wie möglich außen vor bleibt, sondern wir haben ein objektives Kriterium, nämlich das Prädiktionsintervall. Es ist also keine Verschärfung, sondern eine bessere Operationalisierung. Frage ist: Wieso wird das als Verschärfung verstanden?

Moderator Jürgen Windeler: Gibt es jemanden, der sich dazu äußern möchte?

Friedhelm Leverkus: Der Punkt, der uns ein bisschen sorgt, ist: Erst einmal ist es gut, dass ihr versucht, das stärker zu operationalisieren, mehr Guidelines zu geben. Das ist sicherlich ein guter Punkt. Aber wir wissen im Prinzip nicht, welche Konsequenzen das in konkreten Situationen hat. Von daher wäre es wünschenswert, wenn ihr das einführen wollt, dass ihr Simulationsstudien durchführen lasst, um mehr über die Eigenschaften, die das Verfahren hat, zu wissen.

Ralf Bender: Wie sollte so eine Simulationsstudie aussehen? Wir haben jetzt hier besser operationalisiert, wann wir von mäßiger und wann wir von deutlicher Gleichgerichtetheit sprechen. Im Moment kann man alles Mögliche simulieren, aber die Kernfragestellung ist mir da ehrlich gesagt etwas unklar. Was soll man da simulieren, mit welchem Ergebnis?

Friedhelm Leverkus: Was man ja machen könnte, wenn man eine konkrete Studiensituation hat, ist, zu überprüfen, welchen Einfluss das auf diese Beurteilung hat. Man definiert Heterogenitätsgrade und schaut dann, wo man im Prinzip landet. Das würde allen, die davon betroffen sind, ein bisschen mehr Sicherheit geben.

Dieter Hauschke: Ich bin ein bisschen überrascht, teilweise sogar konsterniert. Es ist keine Verschärfung. Es ist ein Tool da, was in einer bestimmten Situation einen weiteren Weg öffnet. Simulationsstudie finde ich auch nicht so gut. Ich sage immer: Beweis durch vollständige Simulation. Die Methodik ist publiziert. Die kann man nachlesen. Deswegen ist das doch schön, dass man so etwas hat.

Guido Skipka: Friedhelm, ich kann dich ein Stück weit verstehen. Die Methode ist publiziert. Wir verwenden sie natürlich, um die Beleglage abzuleiten. Da kann man sich schon fragen, welche Konsequenzen das hat.

Simulationen stelle ich mir auch erst einmal schwierig vor. Nichtsdestotrotz - das habe ich in meinem stillen Kämmerlein schon einmal gemacht - kann man sich natürlich einzelne Konstellationen aufzeichnen, ein bisschen verschieben und gucken, welche Auswirkungen das hat. Ohne dass ich das völlig systematisch reproduzieren kann: Es gibt Situationen, wo man im Vergleich zu vorher vielleicht doch zu einer Aufwertung und gegebenenfalls zu einer Abwertung kommt. Aber man wird in der allergrößten Mehrheit der Fälle zu gar keiner anderen Aussage kommen. Kurz gesagt: Ich habe keine Richtung erkennen können, weder Verschärfung noch Liberalisierung.

Friedhelm Leverkus: Das sind doch tolle Ergebnisse. Es wäre doch schön, wenn du das veröffentlichst, sodass man es sehen kann.

Stefan Lange: Aber lieber Herr Leverkus, warum machen Sie denn nicht selber Simulationen? Warum sagen Sie uns nicht, wie es besser geht? Machen Sie es doch einfach. Ein Beispiel: Ich erinnere mich noch - es ist ein paar Jahre her -, da gab es eine FDA-Guideline zu Nichtunterlegenheitsgrenzen bei Antiinfektiva. Dann gab es daraufhin wunderbare Publikationen, die gezeigt haben, dass diese Guideline nicht der letzte Brüller ist. Machen Sie es doch! Daran hindert Sie doch keiner! Die FDA hat daraufhin ihre Guideline geändert. Wir haben ja vorhin schon besprochen: Nein, wir sind nicht Betonfraktion. Wenn man uns überzeugende Argumente liefert, dann werden wir uns denen natürlich nicht verschließen.

Michael Hennig: Ich habe eine ganz pragmatische Frage zur Anwendung von Prädiktionsintervallen, die Implementierung in der Software. Da würde mich interessieren, mit welchen Softwaretools das IQWiG da arbeitet, ob Empfehlungen ausgesprochen werden können, in welcher Software das implementiert ist. Wir haben es mal recherchiert und haben das noch in sehr wenigen Softwaretools implementiert gefunden. Könnten Sie dazu mal Stellung nehmen?

Guido Skipka: Können wir. Wir konkret rechnen mit SAS diese Dinge aus. Die Formel ist nicht so schwierig. Wir haben sie selber programmiert. Das geht nicht über den Schwierigkeitsgrad von der Berechnung eines Konfidenzintervalls hinaus. Das bekommt man also hin. Wenn Sie aber nach Software fragen, wo das implementiert ist: Es gibt ein R Package MetaReg. Ich bin nicht ganz sicher. Es gibt aber ein neues R Package, wo Prädiktionsintervalle ausgegeben werden.

Stefan Schinzel: Ich möchte noch ergänzen, dass auch in der aktuellen Version von MIX 2.0 Prädiktionsintervalle dargestellt werden können, allerdings nur in Verbindung mit dem Punktschätzer.

Guido Skipka: Es ist vollkommen klar, dass wir bei der Berechnung auch von Konfidenzintervallen Punktschätzer nicht ignorieren können. Sie liefern uns immerhin die Lage des Konfidenzintervalls.

Ralf Bender: Eine kurze Ergänzung zur Software: Das ist im Plan für den Review-Manager enthalten, der Cochrane Collaboration, für die nächste Version, die im nächsten Jahr vorgesehen ist. Da wird es auch Prädiktionsintervalle geben.

Moderator Jürgen Windeler: Ich sehe keine Wortmeldungen zu diesem TOP mehr. Wir machen jetzt eine Mittagspause bis 13:40 Uhr, sodass wir zuverlässig um 13:45 Uhr wieder anfangen können. Draußen steht ein Imbiss, was Warmes und was Kühles. Bedienen Sie sich. Ich wünsche Ihnen guten Appetit und ein bisschen Entspannung.

(Unterbrechung von 13 Uhr bis 13:42 Uhr)

Moderator Jürgen Windeler: Ich schlage vor, obwohl wir noch nicht zu 100 % vollzählig sind, fortzufahren. Ich hoffe, dass Sie gut gestärkt, nicht zu müde sind, aber so opulent war das Essen ja nicht.

Wir kommen zum

Tagesordnungspunkt 5:

Evidenzaufwertung durch indirekte Vergleiche

Ralf Bender: Dieses Thema klang ja schon bei TOP 3 an. Es gibt hier ganz konkret eine Stellungnahme von Sanofi, wo eben die Frage aufgeworfen wird, wie das Institut dazu steht, ob man eben zum Beispiel in dem Fall, wo nur eine einzelne Studie vorliegt, nicht die Evidenzlage durch Hinzunahme indirekter Vergleiche aufwerten könne. Diese grundsätzliche Möglichkeit wurde ja vorhin schon zugesichert, dass man natürlich Evidenzaussagen durch indirekte Vergleiche stärken kann. Die Frage ist, ob eine solche allgemeine Aussage das ist, was man hier hören möchte, oder ob die Beschreibung von konkreten Situationen hier notwendig ist, wann genau wie welche Situation wie aufgewertet wird. Wenn das der Fall ist, dann Bitte von den Stellungnehmenden, solche Situationen zu beschreiben. Insbesondere da habe ich etwas Probleme. Hier wird ganz konkret vorgeschlagen, die Tabelle 2 in unserem Methodenentwurf um diesen Punkt zu ergänzen. Da sehe ich etwas Schwierigkeiten. Man kann diese Möglichkeit sicherlich im Text erwähnen, aber das in Tabelle 2 unterzubringen, da sehe ich einfach von der Struktur der Tabelle 2 her praktische Schwierigkeiten, das umzusetzen. Da sehe ich im Moment nicht, wie das funktionieren kann. Vielleicht haben aber die Stellungnehmenden dazu Vorschläge.

Stefan Schinzel: Wir haben eine ganz konkrete Situation vorliegen. Wir haben in der MS ein Präparat. Das ist belegt durch eine Zulassungsstudie gegen die zweckmäßige Vergleichstherapie, sodass wir da erst einmal per se auf den Hinweis limitiert sind. Wir haben zwei deutlich größere placebokontrollierte Studien, und wir würden diese placebokontrollierten Studien gerne über placebokontrollierte Studien mit dem zweckmäßigen Komparator zu dem indirekten Vergleich kombinieren und das Ergebnis des direkten Vergleichs dadurch stützen. Für uns ist im Grunde genommen die Frage, auch in Anbetracht dessen, wenn man sich anschaut, welche Wertschätzung bisher indirekte Vergleiche erfahren haben, was da im günstigsten Falle zu holen ist. Wäre die Aufwertung auf einen Beleg

vorstellbar, oder sagen Sie generell, dafür gibt es nur einen halben Punkt, wie Herr Dr. Lange es vorhin ausgeführt hat? Falls ja, dieser halbe Punkt ist ja irgendwie aus Kategorien in dieser Tabelle nicht abgebildet.

Ralf Bender: Zu konkreten Projekten kann ich natürlich jetzt nicht sagen, in der Situation…

Stefan Schinzel: Das war nur ein Beispiel.

Ralf Bender: …steigt man von Anhaltspunkt auf Hinweis. Das geht natürlich so nicht. Wir wollen ja hier allgemeine Dinge diskutieren und nicht konkrete Projekte. Welche generellen Anforderungen wir an indirekte Vergleiche stellen, das war zumindest im Groben schon in der Methodenversion 4.0 dargelegt. Wir möchten, dass die grundlegenden Annahmen überprüfbar sind, überprüft werden und dann auch das gewünschte Ergebnis liefern, um aus indirekten Vergleichen zumindest einen Anhaltspunkt oder sogar Hinweis abzuleiten. Das sind die drei Annahmen: Ähnlichkeit, Homogenität, Konsistenz. Die müssen überprüfbar sein. Daraus ergeben sich eigentlich die Anforderungen an die indirekten Vergleiche bzw. an das Netzwerk. Ganz konkret: Bei einem einfachen adjustierten indirekten Vergleich nach Bucher ist die Konsistenz nicht prüfbar. Von daher lässt sich damit nicht großartig etwas aufwerten. Das müssen schon indirekte Vergleiche sein, wo alle drei Annahmen prüfbar sind, geprüft werden und das entsprechende Ergebnis liefern.

Guido Skipka: Eine Ergänzung: Wir diskutieren diesen Punkt intern natürlich auch. Sie hatten vorhin Tabelle 2 angesprochen. Da ist ein zentraler Punkt, wie man zu Belegen, Hinweisen, Anhaltspunkten kommt, natürlich die qualitative Ergebnissicherheit. Da haben wir im Entwurf drei Stufen: gering, mäßig und hoch. Die Frage ist halt jetzt: Wo würde man jetzt einen indirekten Vergleich da einsortieren? Da kann man schon ein bisschen an unserem Entwurf erkennen, wo die Reise hingeht, salopp ausgedrückt. Meine persönliche Meinung ist, dass indirekte Vergleiche natürlich nicht das Evidenzniveau haben wie ein RCT, eher bei geringer, wenn es sehr gut gemacht ist, vielleicht bei mäßiger Ergebnissicherheit landet.

Moderator Jürgen Windeler: Wie weit gibt es noch Wortmeldungen?

Ch.-Markos Dintsios: Die Frage ist, ob es eine Art additive Ausprägung der qualitativen Ergebnissicherheit geben kann, also ob aus den vorliegenden direkten Vergleichen in der Annahme, dass man bei einem Hinweis gelandet ist, ein indirekter Vergleich, der den direkten Vergleich mit einbeziehen würde - das wäre eine Konstellation, so wie ich es verstanden habe -, in dem Fall auch das Dreieckchen schön schließen würde und auch eine Aussage zur Konsistenz erlauben würde. Damit sind zumindest unter der Annahme der Ähnlichkeit zwei der drei Kriterien, die Ralf Bender vorhin genannt hat, erfüllt und, wenn wir Glück haben, vielleicht auch das dritte. Ob das kumulativ sozusagen über einen doppelten Hinweis hin zu einem Beleg führen kann oder nicht - so, glaube ich, war die Frage zu verstehen. Wenn nicht, dann kann man mich von Sanofi-Seite korrigieren.

Stefan Schinzel: Das war im Grunde genommen genau der Inhalt der Frage. Vielleicht noch ergänzend: Es ist ja schon so, dass man durch die Hinzunahme eines indirekten Vergleichs zu einem direkten Vergleich die Präzision der Schätzung des Effektes schon verbessern kann.

Guido Skipka: Da stimme ich Ihnen uneingeschränkt zu.

Ich weiß nicht, ob ich dich, Markos, richtig verstanden habe. Ich denke schon, dass, wenn man einen direkten Vergleich hat - wir haben vorhin über Reproduzierbarkeit gesprochen, die uns wichtig ist - und man nimmt über einen Brückenkomparator noch einen anderen Weg, um die Evidenz aufzufüllen, das auch unter Umständen eine Form ist, um Reproduzierbarkeit zu zeigen. Zumindest ist das meine Meinung.

Stefan Lange: Würden wahrscheinlich die meisten hier am Tisch so sehen. Das große Problem wird eben nur sein in der Tat im Einzelfall, wie es mit der Ähnlichkeit ist. Wenn wir zum Beispiel Studien haben, wo der Komparator best support of care ist, dann wird es schon schwierig, darüber indirekte Vergleiche oder Netzwerke zu konstruieren. Wie kann ich jetzt nachweisen aus Studien, die ich unter Umständen selber gar nicht durchgeführt habe, dass mein Komparator im Grunde genommen dem entspricht? Das wird im Einzelfall vielleicht ein Problem sein. Das sollte man an der Stelle nicht verschweigen.

Alles andere ist unkritisch. Darüber kann man sich sehr schnell einigen. Wann ist eine ausreichende Homogenität da in der Meta-Analyse? Wann ist ausreichende Konsistenz da? Klar, das ist im Augenblick noch nicht so ganz methodisch abgesichert, aber ich glaube, das sollte nicht so ein Riesenproblem sein, aber diese Ähnlichkeitsgeschichte, das wird die Nagelprobe sein.

Guido Skipka: Was ich noch ergänzen möchte, ist - das hat Herr Thomas Kaiser eben schon anklingen lassen -: Man muss natürlich aufpassen, dass man da nicht sehr selektiv vorgeht. Da, wo man es gerne haben möchte, da bemüht man sich, über indirekte Vergleiche noch die Evidenz zu stärken, und da, wo es eher negativ ausfallen könnte, da lässt man das. Irgendwie muss man da schon einen Weg finden, dass diese Selektion nicht stattfindet. Wenn wir sagen, Evidenz kann aufgewertet werden, ja, beim Nutzenparameter ist das so. Wenn wir aber über den Schadensparameter reden, heißt das gleichbedeutend, dass natürlich auch eine Abwertung möglich sein muss.

Thomas Kaiser: Das gilt nicht nur für den Schadensparameter, sondern auch für die Nutzenparameter. Wichtig ist natürlich, dass Sie da eine vollständige und systematische Aufbereitung machen, also nicht sagen: „Für diese zwei Endpunkte haben wir eine gewisse Unsicherheit, deswegen wollen wir uns das anschauen“, sondern die Evidenz als solche ist dann natürlich für die gesamte Fragestellung aufzubereiten.

Markos hat gerade von additiver Ergebnissicherheit gesprochen. Es gibt natürlich gegebenenfalls auch eine subtraktive Ergebnissicherheit. Also Sie haben einen Anhaltspunkt aus direkten Vergleichen und einen Anhaltspunkt für einen geringeren Nutzen aus indirekten Vergleichen. Damit haben Sie möglicherweise gar keinen Zusatznutzen mehr. Insofern würde ich auch sagen: Man kann das als einen weiteren Evidenzkörper betrachten. Das ist aber keine einseitige Richtung, ich schaffe es sozusagen immer, aus einem Hinweis einen Beleg darzustellen, sondern es muss vollständig systematisch gemacht werden, und dann wird man sich anschauen müssen, was im Gesamtergebnis dabei herauskommt.

Ein Hinweis noch: Indirekte Vergleiche - das klang gerade schon bei Stefan Lange an - haben meistens das Problem, dass Sie das gar nicht vollständig auf Basis Ihrer eigenen Studien durchführen, sondern dass Sie da auf andere Studien angewiesen sind. Da haben Sie natürlich auch das große Problem von Publikationsbias, Fehlpublikationen gegebenenfalls. Das ist

sicherlich auch noch einmal ein Appell an alle, eher anders mit diesen Dingen umzugehen, denn das holt einen in den indirekten Vergleichen ein.

Bernhard Wörmann: Aus medizinischer Sicht liegt, glaube ich, der Unterschied am größten darin, dass bei einer Population von Patienten, wo ich gegen einen best support of care vergleiche, gegen eine Population, wo ich gegen eine wirksame Therapie vergleiche, die Krankheitsstadien nicht dieselben sind. Ich glaube, das ist kritisch zu betrachten. Das trifft vielleicht auf Multiple Sklerose zu. In der Onkologie wäre das so, dass es so ist. Best support of care ist Lastline-Therapie. Das sind nicht dieselben und ist auch für die Nutzenbewertung nicht dasselbe wie in der Therapie in einem früheren Stadium. Ich glaube, es ist medizinisch schwierig mit indirekten Vergleichen. Das kann man am besten bei Apixaban sehen. Apixaban ist in einer Studie verglichen worden mit Vitamin-K-Antagonisten. Und die, die die nicht vertragen konnten, wurden gegen ASS verglichen. Das sind nicht dieselben Patienten.

Stefan Schinzel: Ich bin mit Ihnen völlig d'accord, dass natürlich umfassend nach aller Literatur und Studien gesucht werden muss, die für so einen indirekten Vergleich überhaupt infrage kommen. Ich bin mit Ihnen völlig d'accord, dass man natürlich dann alles, was an Endpunkten in diesem Publikationsmanuskript berücksichtigt ist, sich anschauen muss, auch wenn einem die Ergebnisse vielleicht auf das eigene Präparat in dem einen oder anderen Endpunkt nicht so gefallen. Was natürlich ein bisschen ein Problem darstellt, ist, wenn man genügend Endpunkte betrachten kann, sowohl wirksamkeits- wie auch sicherheitsrelevante Endpunkte, wird man immer Endpunkte finden, wo man eine gute Konsistenz zu dem direkten Vergleich findet, und man wird andere Endpunkte finden, wo man das Gefühl hat, das passt nicht so gut zusammen. Vielleicht können Sie noch ein paar Sätze darüber verlieren, wie mit so einer Situation aus Ihrer Sicht umgegangen werden sollte.

Ralf Bender: Diese Bewertung wird zunächst endpunktbezogen gemacht genauso wie Heterogenitätsuntersuchungen in paarweisen Meta-Analysen. Und wenn das für einen Endpunkt homogen ist, dann macht man die Meta-Analyse. Und wenn das für den nächsten zu heterogen ist, dann wird der gepoolte Schätzer nicht gebildet. So ist das bei indirekten Vergleichen auch. Wo es passt, da kann man das Netzwerk rechnen. Und bei dem Endpunkt, wo es nicht passt, da kann man das aus den dargelegten Gründen eben nicht sinnvoll berechnen.

Stefan Lange: Eine Ergänzung: Natürlich bedeutet das wieder eine Erhöhung des Unsicherheitsgrades. Wir sprechen ja über die Möglichkeit, größere Sicherheit zu schaffen - so habe ich es verstanden -, also indirekte Vergleiche als Hilfsmittel, um sozusagen die fehlende zweite Stufe zu bilden. Wenn das aber nicht konsistent funktioniert, funktioniert es dann vielleicht auch im Gesamtergebnis nicht, größere Sicherheit zu bekommen. Das ist dann einfach so.

Als weitere Möglichkeit ist natürlich immer wieder anzuführen, dass sich die Dossierersteller selber Gedanken machen, wie dieses gegebenenfalls zu erklären ist. Und wenn es dafür gute Erklärungen gibt, die das Gesamtergebnis nicht infrage stellen, dann werden wir uns dem sicher auch nicht verschließen, aber es muss halt sinnvoll adressiert und wissenschaftlich begründet werden.

Stefan Schinzel: Ganz kurz: Aus Ihrer Sicht lässt es an der Ähnlichkeit der Studien zweifeln, wenn es einzelne Endpunkte gibt, in der die Konsistenz zum direkten Vergleich nicht so überzeugend ist?

Stefan Lange: Weniger an der Ähnlichkeit der Studien, sondern an der Frage: Erhöhe ich mit der Zunahme des indirekten Vergleichs meine Sicherheit? Da können wieder Zweifel auftauchen, und gegebenenfalls wird es eben nicht gelingen, durch diese Inkonsistenz die Sicherheit zu erhöhen, sondern man wird wieder gucken müssen, ob es etwas mit bestimmten Subgruppen zu tun hat oder auch nicht. Das ist natürlich generisch eine schwierige Diskussion, aber die Studien können durchaus ähnlich genug sein. Es ist dann aber eben nicht so, dass es sozusagen dieses klare Ergebnis gibt, sondern es verbleibt dann die Unsicherheit, die dann eben nicht dazu führen kann, dass man beispielsweise von einem Beweis oder einem Beleg spricht, sondern eben nur auf der Hinweisebene bleiben kann.

Guido Skipka: Noch eine Anmerkung, Herr Schinzel: Sie sagten eben - zumindest habe ich es so verstanden -: Es wäre ja ganz klar, dass, wenn man indirekte Vergleiche macht, man alle möglichen Studien per Recherche systematisch zusammensuchen muss. - Das ist zumindest in der Theorie gar nicht so einfach. Wenn ich mich für einen Vergleich A versus B interessiere, habe da aber keine direkten Vergleiche, suche mir dann zu A versus irgendetwas und B versus irgendetwas die Studien, da finde ich dann noch ein C, D und E, und dann könnte ich aber auch noch nach Studien E versus irgendetwas suchen. Was ich sagen will: Das tritt unter Umständen eine Kaskade los, und Sie suchen sich tot. Die Frage haben wir uns auch schon gestellt: Wie sucht man eigentlich nach den entsprechenden Brückenkomparatoren, und an welcher Stelle macht man dann Schluss?

Bernhard Wörmann: Ist das jetzt eine Einzelberatung?

Moderator Jürgen Windeler: Sie dürfen sich beteiligen oder beraten werden.

Bernhard Wörmann: Dann können wir Schluss machen.

Moderator Jürgen Windeler: Gut, wenn das der Wunsch ist. Gibt es noch Wortmeldungen hierzu? - Ich glaube schon, dass es ein durchaus breites und allgemeininteressierendes Problem ist, was wir hier adressiert haben, jedenfalls was wir an sonstigen Diskussionen außerhalb dieser Veranstaltung oder auch an sonstigen Stellungnahmen bekommen. Dann schließe ich diesen TOP ab.

Wir kommen zum

Tagesordnungspunkt 6:

Methodik für Meta-Analysen mit zufälligen Effekten

Ralf Bender: Hierzu gibt es eine Stellungnahme vom vfa. Der nimmt Bezug auf eine frühere Stellungnahme von Oliver Kuss, der im Auftrag der GMDS und der IBSDR unser Kapitel zu Meta-Analysen kommentiert hat. Damals hat Oliver Kuss darauf hingewiesen, dass es bessere

Methoden gibt als die übliche Standardmethodik für Meta-Analysen mit zufälligen Effekten, so wie sie momentan vom Institut verwendet wird, und das Institut aufgefordert, die vorhandenen besseren Methoden sich anzuschauen und zu vergleichen, um zu einer besseren Methodik zu kommen. Als Erläuterung dazu: Das ist deshalb noch nicht in den Entwurf des Methodenpapiers 4.1 aufgenommen worden, weil gerade die Untersuchung dieser anderen Methoden nicht ganz so einfach ist. Es gibt zahlreiche Methoden, auch verschiedene Kombinationen untereinander. Für diese Untersuchung wurde eigens von der Statistical Methods Group der Cochrane Collaboration eine Arbeitsgruppe eingerichtet, an der ich und auf meinen Vorschlag hin auch Oliver Kuss beteiligt sind, außerdem aus Deutschland Guido Knapp. Mit den Ergebnissen ist frühestens nächsten Jahres zu rechnen. Die Haltung des Instituts ist, nicht voreilig die Methodik zu ändern, sondern auf die Ergebnisse dieser Arbeitsgruppe zu warten. Das ist auch bei der letzten IQWiG-Präsidiumskommissionssitzung der Fachgesellschaften so besprochen worden. Die Fachgesellschaften hatten keine weiteren Einwände, bzw. ich habe Unterstützung wahrgenommen, auf die Ergebnisse der Arbeitsgruppe zu warten. Frage ist, ob das der vfa auch so sieht oder ob er denkt, das ist so wichtig, dass das Institut die Methodik früher ändern sollte.

Weitere Frage: In dieser Stellungnahme werden relativ wahllos eine ganze Reihe von Verfahren genannt. Hat der vfa einen Vorschlag, welches dieser zahlreichen Verfahren in welchen Situationen möglicherweise das beste Ergebnis liefert?

Ch.-Markos Dintsios: Vollkommen d'accord aus vfa-Sicht. Wir kennen natürlich nicht die Gespräche, die zwischen dem Präsidium des IQWiG und den Fachgesellschaften laufen. Wir nehmen daran nicht aktiv teil. Die Aussage, die hier getätigt wurde, führt bei uns zur vollsten Zufriedenheit. Zumindest setzt man sich mit der Materie auseinander. Und wenn es noch nicht so weit gediehen ist, dann kommt das es eben später dran.

Die Nennung der Methoden war nur eine exemplarische, ohne jegliche Präferenz für eine Priorisierung dieser Methoden kundzutun. Wir warten gespannt auf die Ergebnisse der Arbeitsgruppe. Es war nur ein Hinweis, dass die thematisch ja bereits vorher diskutiert wurde und wie man damit umgegangen ist. So ist auch dieser Punkt von uns zu verstehen, nicht einmal vorwurfsweise, sondern eher: Wir wussten nicht, ob man sich dessen angenommen hat oder nicht. Das war es auch.

Moderator Jürgen Windeler: Gut, dann können wir es hier kurz machen. Gibt es Fragen dazu? - Sehe ich nicht.

Ich komme zum

Tagesordnungspunkt 7:

Verschiedenes

Moderator Jürgen Windeler: Gibt es Wortmeldungen aus der Runde?

Ch.-Markos Dintsios: Es hat mit einem Tagesordnungspunkt zu tun, den wir bereits behandelt haben, mit den Konfidenzintervallgrenzen. Ich möchte das nur andiskutieren. Da wurde ja ein Anker festgesetzt in dem Vorschlag, den das IQWiG gemacht hat, bezogen auf die Mortalität, also den Schätzer, der für die Hypothesenverschiebung eingeht, um das Konfidenzintervall herzuleiten. Was ich leider nicht in Erfahrung bringen konnte - glauben Sie mir, ich habe auch Herrn Djulbegovic selber angeschrieben, aber habe es auch indirekt versucht über Akademia, weil ich ja ursprünglich selber mal aus Akademia kam -… Ich habe bis jetzt nie eine Antwort bekommen. Ich weiß nicht, ob Sie diese Studien haben. Ich möchte gerne mal diese zwölf Studien sehen - das sind ja Einzelstudien -, um zu sehen, wie sich die Effektschätzer gestalten und wie die Konfidenzintervalle und diese Effektschätzer aussehen. Denn im Hypothesenshiftvorgehen haben Sie ja die Annahme der zwei pivotalen Studien getroffen. Mich würde es ganz einfach interessieren: Ist denn zumindest IQWiG in den Genuss gekommen, diese zwölf Studien mit dem arbiträren Kriterium - das waren, glaube ich, 2 % der Studien gewesen - relative Effektschätzer jeweils Hazard Ratio kleiner gleich 0,5 jemals zu sehen? Denn die Veröffentlichung hat keinen technischen Anhang. Und es ist auch sehr interessant, auch für die evidenzbasierte Medizinszene, dass einfach eine Riesenanzahl an Studien ausgewählt wurde, ohne jemals die einzelnen Studien richtig zu nennen. Haben Sie da was? Das ist also nicht unbedingt relevant als Methodenvorschlag, sondern relevant hinsichtlich der Transparenz der Informationen, auf der Sie selber ja Ihren Vorschlag aufbauen. Haben Sie die Studien jemals einzeln gesehen? Wenn ja, hätten wir sie auch gerne einmal gesehen.

Guido Skipka: Ich habe die Studien noch nicht gesehen.

Stefan Lange: Aber es gibt Studien, die dieses Kriterium erfüllen.

Monika Nothacker: Im Anschluss daran nur noch eine kurze Sachzusatzfrage, wie Sie genau auf diese Studie gekommen sind.

Stefan Lange: Es ist nicht Ergebnis einer systematischen Recherche, um das direkt vorwegzunehmen. Das sollte man damit nicht überfrachten. Es gibt im Grunde genommen bisher zu dieser spezifischen Frage, wie man Effektstärken in dieser Weise einordnet, im Grunde genommen nichts. Insofern ist sie dem eigenen Erfahrungsschatz an vernünftiger Literatur entsprungen, wo für uns erkennbar das erste Mal jemand eine solche Formulierung vorgenommen hat oder einen Versuch vorgenommen hat, sozusagen einen Durchbruch zu definieren. Ich will nicht behaupten, dass es nicht andere Arbeiten gibt, die das auch gemacht haben. Ich bin aber relativ sicher - das sage ich jetzt mal so; da lehne ich mich ziemlich weit aus dem Fenster -, dass die zu nicht wesentlich anderen Ergebnissen kommen. Ich glaube, dass dieser Anker nach Halbierung des Mortalitätsrisikos eigentlich schon eine ganz vernünftige Definition ist.

Konrad Wink: Ich habe eine Frage, die etwas abgeht von dem, was wir hier besprochen haben. Es war ja hier doch die Biometrie, die die Dominanz hatte. Wir sprechen über Arzneimittel, aber das Wort „Pharmakologie“ kam nie vor. Meine Frage und vielleicht auch ein Vorschlag ist der: Warum berücksichtigen wir an sich nicht pharmakologische Erkenntnisse? Es gibt ja ganz gute Beispiele. Ich meine jetzt nicht die Molekularpharmakologie - da wird sich wahrscheinlich alles ändern; da werden wir wahrscheinlich alle arbeitslos werden; da brauchen wir den IQWiG gar nicht mehr; da lässt

sich schon alles vorher bestimmen -, sondern ich denke an etwas anderes. Wenn Sie das Beispiel Ticagrelor nehmen: Bei Ticagrelor wäre doch eindeutig schon allein von der Pharmakogenetik, von der Pharmadynamik her ein Vorteil erkennbar: Wir haben die rasche Bindung, wir haben die rasche Lösung usw. Es wirkt sofort. Wenn es flutet, kann man es wieder absetzen. Wesentlich besser als bei Clopidogrel und Prasugrel. Das ist ein isolierter Versuch, wo alle Störfaktoren ausgeschlossen sind. Deshalb hat es doch eine gewisse hohe Aussagekraft. Was wir aber hier machen, sind Untersuchungen, die mit vielen Störfaktoren behaftet sind, besonders für den Zusatznutzen, wo man die Kriterien nicht so scharf wie bei der Zulassung formulieren kann, wo viele Störfaktoren das Ergebnis verändern können. Das haben wir da zum Beispiel nicht. Deshalb meine Frage: Warum wird wenigstens der Wirkungsmechanismus nicht einbezogen in die Bewertung der Studien? Der könnte uns eventuell sogar weiterbringen als die Möglichkeit, aus Daten retrospektiv einen Unterschied heraus zu kitzeln. Es mag sich ändern, wenn das vielleicht prospektiv gemacht wird. Dafür sind ja Ansätze da. Aber noch sind wir in dem Stadium, wo fast alles retrospektiv erfolgt mit all den Problemen. Deshalb mein Vorschlag: Warum beziehen wir nicht die Pharmakologie mit ein, zumindest die experimentelle, die uns doch vielleicht rein reduktionistisch saubere Ergebnisse liefern kann?

Moderator Jürgen Windeler: Herr Kaiser, möglichst eine Antwort, ohne dass wir jetzt in Detaildiskussionen zu einzelnen Medikamenten kommen.

Thomas Kaiser: Rasche Bindung, rasche Lösung kann auch in die Irre führen. Entscheidend ist, dass das Herzinfarktrisiko reduziert wird, nicht ob es rasch bindet oder nicht rasch bindet. Wenn Sie sich die Zulassung anschauen, wenn das das entscheidende Kriterium wäre, dann würden solche Studien für die Zulassung reichen. Selbst die Zulassung benötigt solche großen Endpunktstudien, weil man ja aus vielfältigen bisherigen Erfahrungen mit anderen Medikamenten gelernt hat, dass das Verlassen auf die Pharmakologie, also auf diese Mechanismen, häufig in die Irre führen kann, weil diese Medikamente auch an ganz anderen Rezeptoren ansetzen können, andere Wirkmechanismen haben. Für den Patienten ist entscheidend, das Risiko für den Herzinfarkt zu reduzieren, nicht ob es mehr oder weniger bindet.

Konrad Wink: Das wollte ich nicht infrage stellen, sondern mir ging es darum, ob man nicht zusätzlich diese pharmakologischen Untersuchungen hinzuzieht. Es gibt genauso viele Beispiele, wo das gut funktioniert. Ich meine, Ticagrelor war ja gar nicht so schlecht hinsichtlich der Ergebnisse mit den Studien, mit den Endpunkten. Da war vieles konform. Das hat sich bestätigt. Ich könnte Ihnen noch viele andere Beispiele sagen, wo es sich bestätigt hat, aber - das gebe ich zu - auch, wo es sich nicht bestätigt hat, meistens weil wir eben doch nicht genügend wissen. Wenn wir also mehr wissen würden über die Pharmakologie, könnte das Ergebnis auch besser sein und die Übereinstimmung auch.

Thomas Kaiser: Da können Sie sich jetzt zwei Szenarien vorstellen. Das eine ist, bei Ticagrelor kommt dann ein positives Ergebnis in den Endpunktstudien heraus oder es kommt kein positives heraus. Jetzt haben Sie von der Pharmakologie her eine positive Hypothese. Was machen Sie denn jetzt zum Zusatznutzen? Ich meine, wenn nichts Positives herauskommt, dann ist doch die Aussage zum Zusatznutzen klar: Es gibt keinen Beleg für einen Zusatznutzen, egal ob Sie aus der pharmakologischen Überlegung einen potentiellen Vorteil ableiten oder nicht. Entscheidend ist, ob das Risiko für einen Herzinfarkt bei dem

Patienten reduziert wird, nicht ob das Arzneimittel mehr bindet oder weniger bindet. Das heißt, das hat für die Gesamtbetrachtung zum Zusatznutzen gar keine Bedeutung, ob Sie das zusätzlich betrachten oder nicht zusätzlich betrachten. Entscheidend sind die Studien zum Zusatznutzen.

Konrad Wink: Ich möchte Ihnen da ein bisschen widersprechen. In dem Fall ist das gerade nicht der Fall. Sie haben ja im Vergleich zu Clopidogrel und Prasugrel bessere pharmakologische Ergebnisse. Das hat sich nachher auch bestätigt in den klinischen Studien. Hier ging es konform. Da wäre es ja suprativ. Da könnte man es mit einbeziehen.

Thomas Kaiser: Aber wenn Sie das vom Grundsatz her fordern, müssen Sie doch auch die andere Situation durchdenken. Jetzt haben Sie also ein negatives Ergebnis. Würde die Pharmakologie dann plötzlich sagen: „Nein, dann haben wir aber trotzdem einen Zusatznutzen!“ Oder Sie haben es sogar umgekehrt: Sie haben gar keinen Vorteil in der Pharmakologie, aber einen Unterschied in den Studien zum Zusatznutzen. Würden Sie damit die Zusatznutzenaussage aus den Studien infrage stellen? Man kann es ja nicht einseitig sehen. Entweder ist das eine wertvolle Evidenz und sie stellt das infrage oder nicht infrage oder eben nicht. Und wenn es grundsätzlich nur suprativ sein kann, wenn es kongruent ist, dann kann ich auch darauf verzichten.

Konrad Wink: Supratives Ergebnis ist doch wichtig. Das haben wir ja vorhin auch bei den indirekten Vergleichen akzeptiert. Also so schwach würde ich es nicht sehen. Nein, natürlich offen sehen. Ob ich das dann als Zusatznutzen bezeichne noch nicht, nur im Zusammenhang mit den anderen Ergebnissen, aber immerhin: Es wäre für mich ein Anreiz oder eine Motivation, die Sache entweder noch mal weiter zu untersuchen oder noch weitere klinische Studien einzufordern oder einfach noch die Sache weiter zu betreiben. Dagegen, wenn wir von vornherein ein pharmakologisches Ergebnis haben, was unmöglich ist - ich denke zum Beispiel an die COX-2-Hemmer -, da haben wir ja alle gemeint, es seien große Vorteile für die kardiovaskulären Erkrankungen, obwohl es nur die Magenblutung betraf. Da war es ja nicht der Fall. Und da haben wir tatsächlich die Pharmakologen, die gesagt haben, das geht nicht gut, das wird irgendwann zu vermehrten kardiovaskulären Ereignissen führen. Die hatten recht.

Stefan Lange: Ich glaube eher, dass es das Problem ist, es geht halt um unterschiedliche Fragestellungen. Das, was Sie adressieren, ist durchaus interessant, nämlich zum Beispiel der Innovationsgrad eines neuen Arzneimittels. Wir hatten vorhin über Vemurafenib gesprochen. Das war sicher hochinnovativ im Zusammenhang mit zielgerichteten Therapien. Aber das ist nicht das, was wir zu bewerten haben. Wir haben klare Kriterien, die in der Arzneimittelnutzenverordnung niedergeschrieben sind. In der Tat steht da von der Pharmakologie nichts drin, sondern die harten Endpunkte Mortalität, Lebensqualität. Insofern ist es eine interessante Information. Und vielleicht für eine andere Art der Einteilung von Arzneimitteln in Me-too-Präparate oder völlig neue Therapieansätze könnte es eine Rolle spielen, für uns nicht.

Stefan Schinzel: Das führt jetzt wieder zurück zu dem alten Thema. Wir hatten vorhin die Ausführungen gehört, die zu diesem Ankerpunkt für diese Einteilung des erwünschten relativen Risikos geführt haben. Was ich bis heute immer noch nicht richtig verdaut habe, ist im Grunde genommen diese äquidistante Einteilung in Sechstel, die vorgenommen wurde und

die die Schwellenwerte ganz maßgeblich prägt. Wir haben auch in unseren Ausführungen geschrieben, dass im Grunde genommen die Schätzer für das relative Risiko locknormal verteilt sind. Von daher hätten wir eigentlich, wenn schon eine äquidistante Einteilung, dann eine erwartet auf der logarithmischen Skala, also im Grunde genommen eine geometrische Folge. Das ist etwas, was wir immer noch nicht richtig durchdrungen haben.

Guido Skipka: Herr Schinzel, ich danke Ihnen für diese Stellungnahme. Die habe ich mit Interesse gelesen, die mit dieser logarithmischen Skala. Und ich habe mich auch schon hingesetzt und das auseinanderdividiert. Ich habe auch schon einen Würdigungstext dazu. Dummerweise habe ich es gerade nicht parat, weil ich mir das schon vor etlichen Wochen angeschaut habe. Das können wir etwas spannend halten. Ich habe mich damit beschäftigt. Es spielt keine so große Rolle. Aber ich werde - ich hoffe, zu Ihrer Zufriedenheit - den Punkt würdigen.

Stefen Schinzel: Ich fühle mich geehrt, dass ich eine eigene Erwiderung bekomme. Ich erwarte mit großer Spannung Ihre Ausführungen. Danke.

Stefan Lange: Vielleicht anknüpfend an diesen Punkt: Glückerweise, wie Guido Skipka gerade festgestellt hat, wird es wohl nicht so eine Riesenrolle spielen, ob man dem von Ihnen genannten durchaus logischen Argument folgend eine logarithmische Skala macht, was mich aber verleitet, zu sagen: Ich glaube, auch was den Ankerpunkt angeht, dass man das Ganze auch nicht zu hoch hängen darf. Natürlich ist das arbiträr. Das ist uns auch völlig klar. Das ist nicht so, als ob es irgendwie eine gottgegebene Gesetzmäßigkeit dazu gäbe. Das haben wir aber auch nie gesagt. Das gilt aber für jegliche Festlegungen in der Wissenschaft und spezifisch in der Statistik, in der Medizin. An dem Punkt sage ich immer: Das übliche Signifikanzniveau von 5 % hat auch keine Rationale. Ich kenne jedenfalls keine.

Monika Nothacker: Ich habe zwei Fragen. Der Vertreter der DGG ist ja heute nicht hier. Da ging es um die Würdigung von Nutzen und Schaden und die Bewertung, mit welchen Methoden man das macht. Sie haben dort zwei Methoden genannt, nämlich - ich kürze das jetzt mal ab - CA und AHP. Und die QALYs kommen relativ schlecht dabei weg. Vielleicht können Sie noch einmal erläutern, aus welchen Gründen Sie diese beiden Verfahren präferieren. Das ist das eine.

Das andere ist: Sie haben geschrieben, dass Sie sich an der GRADE-Methodik orientieren. Ich hätte gerne, dass Sie ausführen… Bei GRADE ist es ja so, dass am Anfang die Stakeholder sich zusammensetzen und eine Endpunktbewertung durchführen. Wir hatten von der AWMF schon öfters vorgeschlagen, so genannte Scoping-Workshops einzuführen auch für einzelne Verfahren. Und mich würde interessieren - Sie sind gar nicht darauf eingegangen; das hat ja einen Grund -, ob Sie den noch einmal nennen können.

Andreas Gerber: Wir haben uns bewusst dazu entschieden, darauf diesmal nicht einzugehen, weil, wie Sie alle wissen, ein Papier zu AHP ist erschienen, zur Conjoint-Analyse wird es kommen. Und wir denken, dass es Sinn macht. Wir haben alle diese Stellungnahmen mit Interesse gelesen. Da stand auch viel Spannendes drin. Ich habe mir auch schon Entgegnungen, Würdigungen überlegt, aber ich denke, wir sollten… Wir haben diesmal auch eine lange Agenda. Wir werden im Grunde genommen das dann in einer neuen Version unterbringen. Und wir hatten eigentlich hier auch relativ wenige Änderungen. Deswegen

wollen wir die hier auch nicht diskutieren. Die waren eigentlich fast alle nur redaktioneller Art. Die inhaltliche Diskussion sollte man vor dem Hintergrund der beiden Arbeitspapiere führen und dann, wenn es in die Methoden Einzug findet.

Stefan Lange: Ich bin jetzt nicht ganz sicher, auf welche Stelle im Methodenpapier Sie sich beziehen. Wir haben an der Stelle gemeint, was die Bewertung der Ergebnissicherheit angeht, was die Systematik angeht, in der Tat nicht, was einen Scoping-Workshop angeht, wobei ich glaube, dass sich da manches überholt hat. Wir haben heute hauptsächlich über die frühe Nutzenbewertung gesprochen. Ich glaube, wir sind uns alle einig, dass, in den drei Monaten auch noch einen Scoping-Workshop einzurichten, das Ganze etwas überfrachten könnte. Explizit vorgesehen ist er ja auch qua Gesetz für die Kosten/Nutzen-Bewertung und auch für die Punkte, die Sie adressiert haben, was denn geeignete Aggregationsendpunkte sein könnten. Für die sonstigen längeren Bewertungen des IQWiG halten wir nach wie vor unser Vorgehen für vernünftig, zunächst einmal einen eigenen Vorschlag zu machen, der in einen vorläufigen Berichtsplan mündet, der dann aber auch zur Stellungnahme gegeben wird, was letztendlich auf etwas Ähnliches hinauslaufen könnte. Den Stellungnehmenden ist unbenommen, sich zur Wertigkeit von Endpunkten zu äußern, um dann zu einer solchen gewichteten Einteilung zu kommen.

Ch.-Markos Dintsios: Es geht in die gleiche Richtung bezüglich der Gegenüberstellung positiver und negativer Effekte, wo das Methodenpapier des IQWiG selber auch in seiner jetzigen diskutierten 4.1er-Version Teile abändert, auch wenn sie nur redaktionell sind. Es gibt eine Möglichkeit, zumindest, unabhängig von dem festgesetzten Anker für die Mortalität, was den eingehenden Effektschätzer zur Herleitung der Konfidenzintervalle angeht, auch die anderen Endpunkte gegeneinander zu betrachten und gegenüberzustellen, um dann, von mir aus mit demselben Verfahren - das sei dem IQWiG freigehalten - wiederum entsprechende Konfidenzintervalle als Schwellengrößen für die jeweiligen Kategorisierungen herzuleiten. Was mich ein bisschen interessiert, ist, weil genau in diesem Papier sogar Methoden genannt werden, die dazu quasi geeignet wären, wieso, unabhängig von der arbiträren Setzung des einen Ankerpunktes, nicht die Möglichkeit gegeben wird, die anderen Endpunkte, die dort in dieser Tabelle, ursprünglich ja beim Anhang A bei Ticagrelor enthalten sind, gewichten zu lassen. Denn wenn einer der Anker arbiträr ist, dann ist die Sechstelung doppeltarbiträr. Bei dem Anker habe ich ja eine Rationale. Man hat sich bezogen auf die Literaturrecherche und auf das, was aufgefunden wurde. Bei der Sechstelung habe ich überhaupt keine Rationale. Ich kann nur aus der Pragmatik heraus versuchen, sie argumentativ zu unterstützen, aber sie birgt halt eine Gefahr, nämlich dass sie somit ähnlich wie der QALY-Ansatz über alle Indikationsgebiete hinweg dermaßen glättend wirkt, dass sie nicht Spezifika der Indikation, die sich durchaus unterscheiden können, aufgreifen kann. Um es abzukürzen: Mein Vorwurf ist ganz einfach die simplifizierende Vorgehensweise. Die Realität ist kompliziert. Ich verstehe schon aus der Wissenschaftstheorie heraus, dass man einiges vereinfachen muss, aber hier gäbe es sogar die Möglichkeit. Dies zeigt ja sogar das Methodenpapier des IQWiG selber auf.

Moderator Jürgen Windeler: Noch einmal zu der Frage des Hintergrunds der Festlegung: Das IQWiG war in der Situation aufgrund der gesetzlichen und aufgrund der Rechtsverordnungsvorgaben, sich zum Ausmaß äußern zu müssen. Es gab für uns zwei Optionen: Die eine Option war, dass wir uns in einem fröhlichen IQWiG-internen Kreis zusammensetzen und sagen: Was finden wir denn, um eine Mehrheitsentscheidung

herbeizuführen? Dann haben wir uns vorgestellt, was Sie alle dazu sagen würden, besonders zu dem Vorgehen, gar nicht mal zu den Ergebnissen. Dann haben wir uns gedacht, das ist vielleicht kein guter Weg, sondern wir machen einen abstrakten, gut vermittelbaren, in der Tat, Herr Dintsios, einfachen Vorschlag, der sich auch irgendwelcher sehr durchaus berechtigten Wertungen, Bewertungen zwischen verschiedenen Indikationen enthält, sondern wir machen ihn möglichst abstrakt und - das sage ich ausdrücklich und sehr bewusst dazu - möglichst technisch. Denn wir sehen es nicht als unsere Aufgabe an, eine weitreichende inhaltliche Bewertung der Wertigkeit von Endpunkten zu machen über das, was die Rechtsverordnung an Hierarchisierung vorgibt.

Daraus sind die 50 % geworden, und die basieren nicht auf einer umfangreichen systematischen empirischen Analyse, sondern sind eine Festlegung aufgrund des Djulbegovic-Ankers. Deswegen haben wir uns auch die Studien im Hintergrund nicht so wahnsinnig interessiert. Und sie basieren genau aus diesem Grund, aus dieser abstrakten Intention auf dieser Sechstelung. Auch für diese Schiene gilt jetzt wieder ein bisschen das, was den ganzen Tag ein bisschen durchzieht, erstens die Äußerung: Wir haben zwei Jahre lang das jetzt durchzugezogen, immer wieder gemacht, und ich kenne keinen einzigen konkreten Vorschlag, das anders zu machen. Ich kenne verschiedene Äußerungen, es so nicht zu machen, mehr oder wenig spezifisch, aber keinen einzigen Vorschlag, es anders zu machen. Und zweitens: Auch hier bin ich ganz sicher, dass eine Vorgehensweise, wie wir sie jetzt gewählt haben, nämlich ein übergreifendes, eher abstraktes, möglichst wenig wertendes Verfahren zu wählen, zu kritischen Äußerungen führt, wie Sie sie gerade gemacht haben. Dazu muss man doch abwägen und es indikationsspezifisch betrachten. Ich bin sicher, wenn wir ein indikationsspezifisches Vorgehen gewählt hätten, hätten Sie sich darüber beklagt, dass wir das nicht vernünftig operationalisiert und allgemein gemacht haben.

Auch an dieser Stelle wie an allen anderen: Wir sind daran interessiert, dass Vorschläge, die wir machen, oder Vorgehensweisen, für die wir uns entscheiden - auch das Vorgehen bei der Ausmaßfestsetzung ist ein im Stellungnahmeverfahren diskutierter Vorschlag gewesen -… Wir sind daran interessiert, dann zu hören, wie man es besser machen kann. Wir sind interessiert daran, zu hören, wie man es anders machen kann. Wir nehmen zur Kenntnis, aber es interessiert uns nicht besonders, dass man es so auf keinen Fall machen darf. Insofern: Falls jemand einen ganz anderen und guten und umsetzbaren Vorschlag hat, sind wir gerne bereit, den zu diskutieren - sowieso - und uns damit auseinanderzusetzen. Herr Skipka hat gerade in einem anderen Punkt beschrieben, wie er das gemacht hat. Aber ich kenne keinen, wir kennen keinen, bis auf den Hinweis, das müsste man eigentlich indikationsspezifisch machen. Ich bin ganz sicher, indikationsspezifisch wird nicht reichen. Der nächste Schritt wird sein, es dossierspezifisch machen zu sollen, weil ja zwischen verschiedenen Indikationen oder innerhalb einer Indikation zwischen verschiedenen Krankheitsstadien, was auch immer man sich da vorstellen kann, man unterschiedliche Schwellenwerte wählen muss. Man ist bei der Forderung nach einer weitreichenden Beliebigkeit. Dafür haben wir uns jedenfalls bei unserem ersten Aufschlag nicht entscheiden können.

Ch.-Markos Dintsios: Herr Windeler, die Aufgabe war nicht Ohne. Ich glaube, in unserer Stellungnahme - ich habe es auch persönlich oft meinen ehemaligen Kollegen mit auf den Weg gegeben - haben wir ausgeführt, dass es sehr interessant war, mit welchem intellektuellen Ehrgeiz man an diese Thematik herangegangen ist. Letztendlich ist es auch eine Aufgabe, die vom Gesetzgeber über die Selbstverwaltung an das IQWiG weitergereicht

wurde, ohne dass das IQWiG vorher gefragt wurde, ob es sich dem gerne widmen will. Das war ähnlich wie damals die Diskussion zur gesundheitsökonomischen Evaluation.

Nur einem Punkt, wenn Sie mir das erlauben, würde ich gerne widersprechen: Es gab schon Gegenvorschläge. Das, was ich hier gerade zitiert habe, habe ich und haben wir als Verband nicht das erste Mal gemacht. Es gab zwei Workshops im Beisein des IQWiG mit G-BA-Vertretern, mit den Bänken, mit den Fachgesellschaften - Frau Kopp war dabei -, mit dem PEI. Wir hatten einen ameliorierenden, sozusagen einen, diesen Vorschlag, den das IQWiG unterbreitet hat, aufgreifenden Vorschlag unterbreitet, der genau diese eine Schwachstelle bedient, nämlich die Gegenüberstellung der Endpunkte und die Möglichkeit, sie zu gewichten. Das ist ja kein anderer Vorschlag. Er baut sogar auf Ihren Vorschlag auf. Es muss ja versucht werden, das Dilemma, das Sie schildern, was mache ich denn Vollabstraktionsgrad oder ich gehe in die Beliebigkeit hinein, zu lösen. Es kann nicht sein, dass wir, weil wir nicht in die Beliebigkeit hinein wollen, in eine Rigidität kommen, die uns nicht weiterführt. Ich mache ein ganz einfaches plastisches Beispiel: Der Endpunkt Lebensqualität, den die Patientenvertreter immer wieder beim G-BA vortragen, hat in der Onkologie je nach Krankheitsbild, eventuell je nach Stadium eine unterschiedliche Wertigkeit. In der palliativen Versorgung ist das ganz anders in der hinteren Linie als vielleicht vorne. Wenn Sie sagen, wir wissen es nicht - ich gestehe Ihnen gerne dieses Argument zu -, dann ist es an uns, es zu erheben. Dann wissen wir es eben vielleicht hinterher oder vorher. Also in vielen Dingen Übereinstimmung, aber nicht in dem Punkt, es gab keine erweiterten oder keine Alternativvorschläge. Den würde ich so nicht gelten lassen. Aus unserer eigenen Stellungnahme, die wir eingereicht haben, geht dieser ameliorierende Vorschlag definitiv hervor.

Moderator Jürgen Windeler: Zunächst vielen Dank für das Lob. Sie wissen ja auch selber, dass das IQWiG sehr gerne Herausforderungen annimmt, auch auf die es nicht vorbereitet ist, und das dann in der Regel ganz gut auf die Reihe bekommt.

Der zweite Punkt: Ich bleibe bei meiner Formulierung, dass es keinen Vorschlag gab. Eine Äußerung, ihr müsst das irgendwie gewichten und das muss indikationsspezifisch sein, hat mit einem Alternativvorschlag nichts zu tun. Wenn Sie sagen, dass Lebensqualität in verschiedenen Krankheitsphasen unterschiedlich ist, was selbstverständlich nicht nur für die Onkologie gilt, dann ist der nächste Schritt, dass ich gerne wissen möchte, wie Sie das relativ gewichten und mit welchen Grenzen Sie dann arbeiten wollen. Das ist das, was mich interessiert, nicht dass es irgendwie unterschiedlich ist. Das wissen wir auch; da können Sie sicher sein.

Stefan Lange: Der Vorschlag, die logarithmischen Grenzen zu verwenden, wäre ja eine Alternative. Das muss man fairerweise sagen, nur um Herrn Schinzel die Ehre zu gereichen.

Moderator Jürgen Windeler: Bis ich diesen Vorschlag vorhin noch mal in Erinnerung bekommen habe, war meine Äußerung richtig. Jetzt ist sie nicht mehr ganz richtig gewesen.

Stefan Schinzel: Ich habe ja vorhin eine emotionale Reaktion bei Herrn Dr. Lange provoziert, als ich von der normativen Kraft des Faktischen gesprochen habe. Ich habe immer noch Schwierigkeiten, mir vorzustellen, dass Sie aufgrund eines anderen Vorschlages ein anderes Vorgehen festlegen zur Festlegung des Ausmaßes von Zusatznutzen. Und die ganzen

pharmazeutischen Hersteller, deren Dossiers bisher bewertet worden sind, klopfen dann bei Ihnen an und sagen: Jetzt will ich aber auch nach den neuen Kriterien bewertet werden. - Die Frage, die ich in den Raum stellen möchte, ist: Der Zug fährt. Wir haben 50 Dossierbewertungen gesehen. Gäbe es überhaupt eine Chance, im Grunde genommen von dem aktuellen etablierten Vorgehen herunterzukommen? Und welche rückwirkenden Auswirkungen hätte das auf die bereits bewerteten Dossiers?

Thomas Kaiser: Zunächst einmal ist es wirklich ernst gemeint: Es gibt selbstverständlich eine Chance. Wenn Sie gute Argumente für ein anderes Verfahren haben, sind wir dafür offen.

Zur Rückwirkung auf die anderen Verfahren kann man relativ einfach sagen: Weil der G-BA in jedem seiner Beschlüsse geschrieben hat, wie folgen explizit der Methodik des IQWiG nicht, sehe ich da überhaupt kein Problem. Sie müssen auch nicht bei uns anklopfen. Das wird Ihnen wenig helfen. Da können wir nur sagen: Wir sind die falsche Tür. Sie müssen beim G-BA anklopfen. Da wird Ihnen aber wahrscheinlich der G-BA sagen: Weil wir die Methodik nicht verwendet haben, brauchen Sie auch kein neues Dossier einzureichen. Ob mit einem anderen Vorschlag oder überhaupt mit der Diskussion dieses Vorschlags nach Veröffentlichung des Entwurfs und dann Veröffentlichung eines abschließenden Methodenpapiers diese Methodik auch bei G-BA einen anderen Stellenwert bekommt, das muss man abwarten. Man muss ja sehen, dass der G-BA natürlich mit darauf achten muss, wie das mit den Verfahren läuft und dass das hier eine etablierte Methodik im Methodenpapier sein muss. Wir sahen uns von heute auf morgen konfrontiert mit der Situation, haben das deswegen angewandt, haben die Methodik in der ersten Bewertung veröffentlicht und darauf rekurriert. Wir haben jetzt die Diskussion dazu durchgeführt. Da muss man jetzt einfach warten, wie das im Gesamtverfahren auch innerhalb des G-BA aufgenommen wird, inklusive der berücksichtigten Stellungnahmen, der Würdigung des Ganzen.

Aber noch einmal von unserer Seite: Wenn Sie bessere Argumente haben, aber eben nicht nur Argumente, sondern auch einen besseren Vorschlag, dann sind wir selbstverständlich dafür offen, wie an allen anderen Stellen auch.

Stefan Lange: Auch hier wieder der Appell an alle bzw. die, die in den Firmen oder Institutionen arbeiten - vielleicht ist die Zahl von 50 noch nicht ausreichend, aber vielleicht wenn wir mal 100 zusammenhaben, oder man kann sich auch alte EPAs nehmen, keine Ahnung -: Es ist doch interessant, auch hier eine empirische Forschung zu betreiben. Welche Ergebnisse würden denn andere Vorschläge, die vielleicht einen ähnlichen Abstraktionsgrad haben, bringen? Oder ist vielleicht die Verteilung - wir hatten es vorhin kurz angesprochen - der Ergebnisse, so wie sie jetzt sind - man kann natürlich das Minimum, das Maximum der Subgruppen, sonst etwas nehmen, präferenzgewichtet, mit AHP gekreuzt oder mal den QALYs genommen… Wir sind doch alle noch am Lernen und interessiert daran, welche empirischen Ergebnisse das bringt und wie man sie gegebenenfalls noch verbessern kann. Natürlich kann im Einzelfall ein anderes Verfahren oder können andere Grenzwerte dort zu einem günstigeren Ausgang führen. Was wir aber nie wissen, ist, wie es eigentlich in der Gesamtgewichtung aussehen wird, weil natürlich das, was ich auf der einen Seite mache für die Nutzenendpunkte oder die positiven Ergebnisse, gleiche Auswirkungen gegebenenfalls auf der Schadenseite hat, sodass dann in der Gesamtabwägung möglicherweise vielleicht gar

noch so etwas verschiedenes herauskommt. Das ist es, glaube ich, durchaus wert, weiter zu untersuchen. Das ist auch unseren schmalen Schultern zu viel. Insofern hier wieder die Bitte, uns dabei zu unterstützen, das zu untersuchen, welche Eigenschaften ein solches Verfahren hat.

Ch.-Markos Dintsios: Ich möchte nur darauf hinweisen, dass außerhalb der biometrischen Methoden es auch andere Disziplinen gibt, die sich durchaus mit ähnlichen Fragestellungen befassen, nämlich der Entscheidungsfindung. Selbst zwei Verfahren, die in Ihrem eigenen Methodenpapier - ich war ja teilweise an diesen Methoden mit involviert, aktiv - genannt werden, gehören nun mal zu einer anderen Disziplin. Sie kommen aus der Multi-Kriterien-Entscheidungstheorie oder Entscheidungsfindung. Die basiert auf der ökonomischen Theorie. Mein Appell und mein Petitum: Das IQWiG muss sich öffnen, weil es eine Empfehlung ausspricht nach einem gewissen Paragraphen des Sozialgesetzbuchs. Die Biometrie hört irgendwo auf. Deswegen heißt das nicht, dass man den Entscheidungsträger, den man ja bedient als Auftragsinstitut, nicht noch etwas mit auf den Weg geben kann oder sollte.

Was mich ein bisschen missmutig stimmt, ist, dass man aus der Biometrie heraus manchmal Sachen versucht, wie zum Beispiel die Klassifizierung des Zusatznutzens, was ja nicht eine originäre Frage der Biometrie war. Denn bis dato hat das IQWiG selber eher dichotom das Ganze nach außen hingebracht, Zusatznutzen ja, nein. Nach dem AMNOG hat man sich derselben methodischen Kits ein bisschen angepasst, weiterhin bedient.

Ein Appell noch mal: Schaut mal auch links und rechts, was es an anderen Disziplinen gibt. Da gibt es einiges. Selbst im Methodenpapier des IQWiG sind solche Methoden enthalten.

Das noch als Antwort an Herrn Windeler persönlich, weil er gemeint hat, es gäbe keine Verbesserungsvorschläge: Na ja, man kann diese Gewichte zum Beispiel selbst über eine Conjoint-Analyse erheben. Als Beispiel. Sie haben die Antwort auf Ihre Frage schon im eigenen Methodenpapier drin, nur lösen Sie sich bei solchen Fragestellungen von der Biometrie. Ob Sie wollen oder nicht, die evidenzbasierte Medizin kann nicht auf alles eine Antwort geben. Schön wäre es, dann wäre es eine Panakeia. Ist es aber noch nicht.

Moderator Jürgen Windeler: Herr Dintsios, ist muss gestehen, dass ich das alles schwer übereinander bringe. Wenn Sie sagen, dass in unserem Methodenpapier Methoden stehen, die nicht Biometrie sind, dann ist doch Ihre Aussage, dass das alles nur Biometrie ist, offensichtlich falsch.

Der zweite Punkt ist - Herr Gerber hat schon darauf hingewiesen -, dass wir natürlich darüber nachdenken, diese beiden Verfahren, die dort auch drinstehen - andere stehen dort auch drin; es stehen im Methodenpapier sogar qualitative Methoden, wie Sie wissen -, in die frühe Nutzenbewertung zu integrieren. Habe ich Ihr Votum jetzt so verstanden, dass wir dafür sorgen sollen, dass in den Modulen sich Ergebnisse, das heißt natürlich nach IQWiG-Vorstellungen valide Ergebnisse etwa zu AHP und Conjoint-Analyse finden sollten, die selbstverständlich die pharmazeutischen Unternehmer dort hineinschreiben müssten, damit eine valide Nutzenbewertung auch unter Aspekt dieser entscheidungsorientierten Verfahren gemacht werden kann?

Ch.-Markos Dintsios: Ich habe gesagt, dass sich das IQWiG auf die Biometrie konzentriert, was auch vollkommen nachvollziehbar ist, aber dass einige Fragen, die dann dem Entscheidungsträger als Empfehlungen weitergegeben werden, nicht alleine mit biometrischen Methoden beantwortet werden können. Sie hatten eingangs gesagt: Wie sollte man das machen? Sie sehen nichts anderes. Das war ihr vorletztes Statement. Und ich habe Ihnen gesagt: In Ihrem eigenen Methodenpapier gibt es zwei Verfahren. Die kommen aus der Entscheidungstheorie, also heißt das nicht, dass Sie ausschließlich in den Methoden nur Biometrie enthalten, aber Sie kombinieren die Methoden nicht. Sie bleiben bei der Sechstelung als Effektschätzer, der eingeht, um im Rahmen der Hypothesenverschiebung die entsprechenden oberen Konfidenzintervalle herzuleiten. Mein Petitum war: Bedienen Sie sich doch auch den anderen Methoden, die in Ihrem Methodenpapier enthalten sind. Ich bin nicht jemand, der nicht valide Ergebnisse propagiert. Den Schuh würde ich mir nicht anziehen, auch wenn Sie mir den unterschieben wollen, Herr Windeler. Dazu bin ich auch aus derselben Disziplin hervorgekommen.

Bernhard Wörmann: Stören wir Sie sehr bei Ihrem Zwiegespräch?

Ch.-Markos Dintsios: Weiß ich nicht, Herr Wörmann. Das müssen Sie Herrn Windeler fragen.

Moderator Jürgen Windeler: Es gab gerade einen interessanten Versuch einer Klärung. Aber wenn es die anderen stört, dann bin ich gerne bereit, diese Klärung bilateral zurückzustellen.

Monika Nothacker: Interessanterweise wollte ich vielleicht in die ähnliche Richtung, aber dann doch anders als Herr Dintsios. Wir hatten heute zwei Punkte, relatives Risiko und Wichtung, Wertung von Endpunkten, Gegenüberstellung, wo Sie beide ja sehr aktiv geworben haben um bessere Vorschläge, wo vor allem beim letzten dieser Ankerpunkt nicht auf eine systematischen Bewertung basierte. Ich wollte eigentlich noch mal werben für den Austausch und das Einbeziehen des Wissens von anderen Fachgesellschaften, die sich zum Beispiel mit Lebensqualität beschäftigen, oder die Biometriker. Herr Köbberling sagte, Fachgesellschaften sind sehr interessengetrieben. Aber es gibt auch sehr hohe Expertise darin. Das wollte ich nur noch mal unterstreichen, dass man diesen Austausch auch pflegt.

Moderator Jürgen Windeler: Danke.

Dieter Hauschke: Ich möchte ergänzen insbesondere zur Rolle der Biometrie, was ich gerade gehört habe. Ich kann das nur unterstützen, dass die Fachgesellschaften mit einbezogen werden sollen, das heißt, die GMDS oder die Biometrische Gesellschaft. Ich habe ein bisschen den Eindruck, dass sehr viele Leute in vielen Pharmafirmen über biometrische Themen reden. Von den sieben haben wir hier sechs nur biometrische Themen, aber die Fachkompetenz ist nicht bei den Leuten in der Industrie. Das fällt mir wiederholt auf. Ich glaube auch, dass die Industrie erkennen muss, dass Biometriker sich mit diesen Themen beschäftigen sollen. Das wird meiner Meinung nach bislang noch nicht gemacht.

Moderator Jürgen Windeler: Bevor wir jetzt in ganz allgemeine Gefilde geraten, wäre meine Frage, ob noch jemand den dringenden Wunsch verspürt, ein Thema, eine Frage mitgebracht, ohne deren Beantwortung und Aussprache der hier nicht gehen möchte.

Stefan Schinzel: Ein Punkt, der ja auch adressiert wird in dem Entwurf, ist, wie man mit stetigen Zielgrößen umgehen soll. Da wird ja vorgeschlagen, Responsekriterien zu verwenden, damit man wieder als Abstandsmaß das relative Risiko einsetzen kann. Nun gibt es natürlich auch Endpunkte, wo einmal die Frage offen ist, welche klinische Relevanz sie haben, aber auch wo es im Grunde genommen keine etablierten Schwellenwerte gibt. Was zur Methodenkiste des Biostatistikers gehört, ist, die Gruppen zusammenzulegen, Median festzulegen und diesen Median dann als Responsekriterium zu verwenden. Das ist natürlich ein bisschen unbefriedigend, weil das natürlich von Studie zu Studie unter Umständen anders aussieht. Wie stehen Sie zu diesem Splittermedian in der Situation, dass man einen stetigen Endpunkt hat?

Stefan Lange: Ich glaube, der kritischste Punkt bei dieser Frage ist, der sich dann auch für uns in der Bewertung stellt, wenn das eben, wie Sie zurecht sagen, eine gar nicht so seltene Situation ist, dass es keine solchen allgemeinen, anerkannten oder von mir aus validierten oder validen Trennpunkte gibt, dass sie nicht ergebnisgetrieben sind. Insofern würde wieder ein sehr abstraktes Vorgehen, zum Beispiel einer Orientierung an einem Quantil, und zwar an einem üblicherweise verwendeten Quantil - das ist der Median -, sicher bei uns größeres Vertrauen finden, als wenn Sie sich am 37,36-%-Quantil orientieren würden. Als kürzere Antwort: Ja, ich glaube, dass das ein vernünftiges Vorgehen ist, wenn man nichts anderes hat. Wichtig ist, dass es erkennbar nicht ergebnisgetrieben ist. Das ist das Entscheidende, würde ich jetzt so sagen.

Eine andere Möglichkeit wäre natürlich, zu überlegen, wie das über die gesamte Verteilung hinweggeht. Ist es unter Umständen über die gesamte Verteilung hinweg proportional? Dann sind wir sowieso sozusagen bei der automatischen Umrechnung über allgemeinverfügbare Formeln von solchen stetigen in irgendwelche Risikomaße.

Ich glaube, da gibt es verschiedene Möglichkeiten. Aber lange Rede, kurzer Sinn: Median ist sicher ein ganz vernünftiger Trennpunkt.

Moderator Jürgen Windeler: Jetzt gucke ich noch einmal. Ich habe im Moment noch nicht den Eindruck, dass noch jemandem etwas einfällt, auf der Seele brennt. Dann bedanke ich mich an dieser Stelle. Ich beschließe die Erörterung formal. Ich bedanke mich sehr, dass Sie da waren, freue mich, Sie vielleicht auch, dass wir etwas eher fertiggeworden sind, als wir wollten. Ich bedanke mich für Ihre Beiträge, auch Ihre Fragen, auch Ihre kritischen Beiträge, die uns in der Regel wenig ärgern, aber dann doch ein bisschen weiterhelfen.

Wir werden das schriftlich aufarbeiten. Sowohl die heutige Erörterung als auch die entsprechende Würdigung der schriftlichen Stellungnahmen sowie die fertige Überarbeitung des Methodenpapiers werden veröffentlicht werden. Wir werden - das kann ich bereits ankündigen - sozusagen das nächste Paket Anfang 2014 in Marsch setzen, also den Entwurf veröffentlichen, wieder in ein ähnliches Verfahren, was die Stellungnahmen angeht, eintreten wie jetzt.

Dann habe ich, glaube ich, alles gesagt, bis auf den Umstand, dass Sie draußen zur Stärkung für den Heimweg noch Kaffee und Kuchen finden. Vielen Dank und gute Heimreise!

Anhang A – Dokumentation der Stellungnahmen

Inhaltsverzeichnis

A.1 – Stellungnahmen von Organisationen, Institutionen und Firmen

A.1.1 – Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF)

Arbeitsgemeinschaft der
Wissenschaftlichen
Medizinischen
Fachgesellschaften e.V.

Association of the
Scientific
Medical
Societies in Germany

Stellungnahme
der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF)

zum Entwurf des IQWiG–Methodenpapiers:

"Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1" vom 18.04.2013

Die AWMF wurde am 22.04.2013 um ihre Stellungnahme zum oben genannten Vorbericht gebeten. Die AWMF hat ihrerseits ihre thematisch betroffenen Mitgliedsfachgesellschaften gebeten, bei gesehenem Bedarf eine eigene Stellungnahme zu verfassen. Die bis einschließlich 20.05.2013 bei der AWMF eingegangene Stellungnahmen der GMDS/IBS-DR sowie der dggö sind der Stellungnahme der AWMF als Anlage beigefügt. Ebenso beigefügt ist die Stellungnahme des Deutschen Netzwerks für Versorgungsforschung (DNVF), mit dem eine Kooperation seitens der AWMF besteht. Diese Stellungnahmen sind ebenfalls Bestandteil der AWMF-Stellungnahme.

1. Allgemeine Anmerkungen

Die AWMF begrüßt im Sinne der Übersichtlichkeit und Bearbeitbarkeit das Vorliegen eines Dokuments, das ausschließlich die geänderten Abschnitte des Methodenpapiers 4.0 enthält. Wir erwarten, dass das Gesamtdokument 4.1 nach Einpassung der geänderten Abschnitte ebenfalls zur Stellungnahme zugänglich gemacht wird und werden dazu ggf. weitere Anmerkungen machen. Dabei möchten wir auf unsere in früheren Stellungnahmen geäußerten Verbesserungsvorschläge hinweisen, hier insbesondere zur frühzeitigen Einbindung des Sachverstands von Vertretern der Wissenschaftlichen Medizinischen Fachgesellschaften und der Patientenorganisationen in die Verfahrensabläufe des IQWiG ohne produktspezifische Ausnahmen (z.B. Rapid Reports, Dossier-Bewertungen) sowie zur Planung strukturierter Fortschreibungen von als vorläufig anzusehenden Nutzenbewertungen (v.a. Dosierbewertungen)[1].

Zu oben genannten Vorschlägen des IQWiG zur Aktualisierung und Neuaufnahme von Abschnitten geben wir folgende Stellungnahmen ab:

[1] Stellungnahme der AWMF zum Entwurf des Methodenpapiers Version 4 des IQWiG vom 08.03.2011 zum Aspekt: Produktspezifische Verfahrensabläufe. Verfügbar: http://www.awmf.org/fileadmin/user_upload/Stellungnahmen/Medizinische_Versorgung/Stellungnahme_IQWiG-Methoden_IV.pdf

2. Zu den geänderten Abschnitten 2.1.1 „Bericht“ und 2.2.3 2“Review der Produkte des Instituts“

Wesentliche Änderung: Darstellung des externen Reviews für Vorberichte als optionaler Schritt

Die neben der Anhörung zum Vorbericht bislang vorgesehene externe Begutachtung wird nun für alle Produkte des IQWiG als Option eingeschränkt. Die Gründe für diese Änderung werden nicht genannt. Dies zu tun, ist im Sinne von Transparenz und ggf. besserer Nachvollziehbarkeit ebenso wünschenswert wie die Darlegung der Kriterien für die Veranlassung externer Reviewverfahren und deren Umsetzung.

3. Zu dem geänderten Abschnitten 3.1.4 (Neu: „Endpunktbezogene Bewertung“) und 3.1.5 („Zusammenfassende Bewertung“):

Wesentliche Änderung: Konkretisierung der Anforderungen an die Beleglage zur Formulierung von Nutzenaussagen mit unterschiedlichen Aussagesicherheiten

3a. Die Ergänzung einer Abschnittsüberschrift „**Endpunktbezogene Bewertung“** unterstreicht die Bedeutung der Differenzierung der Evidenzbewertung im Hinblick auf relevante Endpunkte, die vor Beginn festzulegen und zu gewichten sind.

Für diese werden Anforderungen konkretisiert, die sich auf das Vorliegen und die Beurteilung der Aussagesicherheit gleichgerichteter Effekte beziehen. Diese Konkretisierung (explizite Unterscheidung zwischen qualitativer und quantitativer Ergebnissicherheit, Ausdifferenzierung der Anforderungen an die Beleglage für die unterschiedlichen Aussagesicherheiten beim Vorliegen von Studien derselben qualitativen Ergebnissicherheit (Tab. 2) sowie Ausführungen zu Studien mit unterschiedlicher Ergebnissicherheit) ist zu begrüßen.

Hinsichtlich der Bewertung eines Studienergebnisses als Nutzenbeleg wird ausgeführt: *„Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und ihre Ergebnisse besondere Anforderungen zu stellen“.* Diese *„besonderen Anforderungen“* sollten durch das IQWiG konkretisiert werden. Die AWMF unterstützt dabei nachdrücklich die Einschätzung des IQWIG, dass aus Einzelstudien in der Regel nur Hinweise abgeleitet werden können.

Für weitere Anmerkungen zu diesem Abschnitt verweisen wir auf die Stellungnahme der GMDS.

3b. Im Abschnitt **„Zusammenfassende Bewertung“** heißt es: *"Eine Möglichkeit der gleichzeitigen Würdigung von Nutzen und Schaden ist die Gegenüberstellung der endpunktbezogenen Nutzen- und Schadenaspekte. Dabei werden die Effekte auf alle Endpunkte (qualitativ oder semiquantitativ) gegeneinander abgewogen mit dem Ziel, zu einer endpunktübergreifenden Aussage zum Nutzen bzw. Zusatznutzen einer Intervention zu kommen."*

Als Methode für die zusammenfassende Endpunktbewertung werden als Alternative zu QALYs kurz zwei Arten von Nutzwertanalysen angedeutet: der „Analytic Hierarchy Process (AHP)“ und die „Conjoint-Analyse (CA)“. Dazu wird angemerkt: „*Eine quantitative Gewichtung* [der Endpunkte] *unter Verwendung von Summenscores oder Indizes sollte prospektiv zum Zeitpunkt der Auswahl der zu untersuchenden Endpunkte erfolgen“*.

Die AWMF begrüßt ausdrücklich, dass das IQWiG die von uns schon sehr früh angemahnte Kosten-Nutzenabwägung (Nettonutzen) weiter konkretisiert und umsetzen will.

Im Methodenpapier sollte jedoch deutlich gemacht werden, inwieweit die Gewichtung von Nutzen- und Schadenkriterien durch das IQWiG erfolgen und wer hierfür einbezogen werden soll. Da das IQWiG sich erklärtermaßen an der GRADE-Methodik orientiert, wäre eine Bewertung der Relevanz der Endpunkte unter Einbeziehung der thematisch betroffenen Vertreter der Wissenschaftlichen Medizinischen Fachgesellschaften und der Patientenorganisationen zu erwarten (z.B. im Rahmen eines Scoping-Workshops).

Bei Verwendung von Scores oder Indizes müssen natürlich auch die Effekte auf die einzelnen Endpunkte transparent gemacht werden, so dass die zusammenfassende Bewertung durch Scores nachvollziehbar und ggf. reproduzierbar ist. Die Ergebnisse der angewandten Methoden zur Zusammenfassung sind dabei nur als Entscheidungshilfen zu verstehen, die die abschließende Beurteilung durch Entscheider unterstützen können aber nicht präjudizieren dürfen. Es wäre darüber hinaus wünschenswert, wenn das IQWiG an dieser Stelle keine fixe Vorentscheidung für Methoden trifft. Die Auswahl der Integrationsmethode wird im Einzelnen von der Indikation und Fragestellung abhängen. Die Limitationen der vom IQWiG präferierten Methoden (AHP und CA) sollten im Methodenpapier benannt werden. Wir empfehlen zudem eine Erprobung der genannten Methoden ex ante und eine Evaluation ex post im Rahmen der Verfahrensabläufe des IQWiG.

Für weitere Anmerkungen verweisen wir auch auf die Stellungnahme der dggö.

4. Zu Änderungen des Abschnitts 3.3.3 und Anhang: Operationalisierung der Feststellung des Ausmaßes des Zusatznutzens sowie dessen Rationale

Die AWMF begrüßt grundsätzlich die Bestrebung des IQWiG, Nutzenbewertungen durch Kriterien zur Feststellung des Ausmaßes des Zusatznutzens nach §5 Abs. 7 der Arzneimittel-Nutzenbewertungsverordnung möglichst vergleichbar zu gestalten.

Insbesondere die Aufnahme und Gewichtung der gesundheitsbezogenen Lebensqualität stellt einen wichtigen Schritt zu einer angemesseneren Nutzenbeurteilung dar. Eine sinnvolle Berücksichtigung der gesundheitsbezogenen Lebensqualität setzt jedoch voraus, dass Verfahren für die Auswahl und Auswertung der verwendeten Instrumente in Abhängigkeit von der Fragestellung konkretisiert werden (siehe auch Stellungnahme des DNVF).

In Abschnitt 3.3.3 wird begründet, warum das IQWiG als Effektmaß zur Beurteilung des Ausmaßes des Zusatznutzens relative Maße gegenüber der Nennung absoluter Effektstärken vorzieht. Die Begründung ist nachvollziehbar, auch wenn anzumerken ist, dass immer auch die zusätzliche Kenntnis absoluter Unterschiede für die Beurteilung der klinischen Relevanz eines Effekts notwendig und wesentlich ist. Dies zeigt sich indirekt in der vom IQWiG geforderten Inzidenz von mind. 5% schwerwiegenden Nebenwirkungen zusätzlich zum geforderten relativen Effektausmaß bei der Einteilung der Effektmaß-Kategorien.

Im Anhang werden Operationalisierungen zur Frage der erforderlichen Effektstärken für die Einordnung einzelner Zielgrößen als erheblich, beträchtlich oder gering getroffen, um *„zunächst die [...] zu treffenden Werteurteile möglichst gering zu halten und diese explizit zu machen.“*

Dabei werden die Schwellenwerte zur Feststellung des Effektausmaßes der Zielgrößenkategorien einer Hierarchie unterworfen:

1. Gesamtmortalität (obere Konfidenzintervallgrenze des erzielten Effekts bei erheblichem oder beträchtlichem Zusatznutzen 0,85/0,95),
2. Schwerwiegende Symptome/Lebensqualität (obere Konfidenzintervallgrenze 0,75/0,90) und
3. Nicht-schwerwiegende Symptome (obere Konfidenzintervallgrenze -/0,80).

Die Konfidenzintervallgrenzen sind aus den für die jeweilige Kategorie zu erzielenden Effekten abgeleitet. Die Grenzwerte wurden ausgehend von einem als „erheblich" gesetzten Effekt von 0,5 für die Gesamtmortalität abgestuft. Hierfür wird auf eine einzige Literaturstelle verwiesen, in der zudem auf den arbiträren Charakter der getroffenen Festlegung explizit verwiesen wird und die lediglich onkologische Fragestellungen betrachtete[2].

Alle in dieser Stellungnahme genannten Fachgesellschaften kommentieren kritisch die starre, nahezu mechanistische Beurteilung der Nutzeneffekte mit fest integrierter Hierarchisierung. Es verwundert insbesondere, dass für diese Fragestellung keine systematische Literatursuche und -bewertung erfolgte bzw. die gewählte Vorgehensweise nicht im Lichte anderer, bereits breiter etablierterer Ansätze zur Einschätzung der Relevanz eines Effektes (z.B. Cohen´s d) diskutiert wird. Der Versuch einer empirischen Annäherung wurde offensichtlich nicht unternommen. So ist anzunehmen, dass die Eignung der Schwellenwerte krankheitsspezifisch bzw. stadienabhängig unterschiedlich ist und diese ggf. entsprechend angepasst werden müssen. Weiterhin müssen auch Konstellationen bedacht werden, bei der die Senkung der Gesamtmortalität keinen patientenrelevanten Endpunkt darstellt, da die Erkrankung, resp. ihre Therapie nicht lebensbedrohlich ist.

Darüber hinaus stellt die Feststellung der Relevanz eines Effektes grundsätzlich ein auf die konkrete Situation zu beziehendes Werteurteil dar. Sowohl die Gewichtung von Endpunkten als auch die Beurteilung von Effektstärken können sich zum Beispiel vom Übergang einer kurativen in eine palliative therapeutische Intention radikal ändern.

Eine Erprobung und anschließende Evaluation der genannten Konzepte des IQWiG erscheint dringend erforderlich und ist entsprechend im Methodenpapier vorzusehen. Zumindest die Anerkennung begründeter Ausnahmen sollte von Anfang an in das Methodenpapier aufgenommen werden. Schlussendlich ist auch hier festzuhalten, dass die Methoden des IQWiG als Entscheidungshilfen zu verstehen sind, die eine abschließende Beurteilung durch die Entscheidungsträger unterstützen, aber selbst nicht präjudizieren dürfen.

5. Zu Änderungen des Abschnitts 7.3.8: Metaanalysen

Wesentliche Änderung: Verwendung von Prädiktionsintervallen für Meta-Analysen mit zufälligen Effekten

Wir verweisen hierzu auf die Stellungnahme der GMDS.

---

[2] Djulbegovic B, Kumar A, Soares HP, Hozo I, Bepler G, Clarke M, Bennett CL.
Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006. Arch Intern Med. 2008 Mar 24;168(6):632-42.
Zitat:" We determined the proportion of discoveries that were "breakthrough interventions." This was arbitrarily defined as interventions judged by the original researchers to be so beneficial that they should immediately become the new standard of care or that had an effect size so large that they reduced the death rate by 50% or more (ie, the HR for death was 0.5 or less)".

Zur Diskussion und für eventuelle Rückfragen zu unseren Kommentaren stehen wir gern zur Verfügung.

Düsseldorf, 21. Mai 2013

Ansprechpartner/Kontakt:
Dr. med. Monika Nothacker, MPH, [redacted]
Prof. Dr. med. Ina Kopp, [redacted]
Prof. Dr. Hans-Konrad Selbmann, [redacted]

AWMF
Ubierstr. 20
40223 Düsseldorf

Anlagen:

- Gemeinsame Stellungnahme der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS) und der Deutschen Region der Internationalen Biometrischen Gesellschaft (IBG-DR) vom 15.05.2013
- Stellungnahme des Deutschen Netzwerks Versorgungsforschung (DNVF) vom 16.05.2013
- Stellungnahme der Deutschen Gesellschaft für Gesundheitsökonomie (dggö) vom 20.05.2013

A.1.2 – Arzneimittelkommission der deutschen Ärzteschaft (AkdÄ)

Arzneimittelkommission der deutschen Ärzteschaft
Fachausschuss der Bundesärztekammer

Stellungnahme der Arzneimittelkommission der deutschen Ärzteschaft (AkdÄ) zur

„Aktualisierung einiger Abschnitte der allgemeinen Methoden Version 4.0 sowie neuer Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“
des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Entwurf vom 18.04.2013

Berlin, den 22.05.2013

www.akdae.de

Ad 3.1.4 Endpunktbezogene Bewertung

In dem neugefassten Abschnitt 3.1.4 „Endpunktbezogene Bewertung“ ist das Vorgehen bei der Ableitung der Belege für die Situationen präzisiert und erweitert worden, in denen wegen Heterogenität der Ergebnisse der Einzelstudien kein gemeinsamer Effektschätzer berechnet werden kann. In diesen Situationen kann es in speziellen Fällen sinnvoll und gerechtfertigt sein, Aussagen zur Beleglage von Effekten zu machen, auch wenn diese nicht quantifizierbar sind – als (extremes) Beispiel: alle Studien des Pools zeigen signifikante Effekte der gleichen Richtung, die Effektschätzer variieren aber so stark, dass Heterogenität vorliegt und keine quantitative Metaanalyse möglich ist.

Nach dem Methodenpapier 4.0 können bisher in diesen Situationen gegebenenfalls auch Belege abgeleitet werden, wenn bestimmte Kriterien erfüllt und so genannte „gleichgerichtete Effekte“ nachweisbar sind. Dem Nachweis „gleichgerichteter Effekte“ wird bei der Bestimmung der Beleglage dann gleiche Bedeutung beigemessen wie einem signifikanten Ergebnis einer Metaanalyse bei homogenem Studienpool. Die Überarbeitung des Methodenpapiers sieht vor, die Kriterien für „gleichgerichtete Effekte“ zu präzisieren und zudem die Unterkategorien „deutlich gleichgerichtet“ und „mäßig gleichgerichtet“ einzuführen.

Zum Nachweis „gleichgerichteter Effekte“ soll jetzt zunächst das so genannte Prädiktionsintervall betrachtet werden. Das Modell des Prädiktionsintervalls wurde entwickelt, um Ergebnisse von „Random-Effects“-Metaanalysen für praktische Umsetzungen und Entscheidungen besser interpretierbar zu gestalten, vor allem bei Vorliegen von Heterogenität (1;2). Es kann vereinfacht als der Größenbereich bezeichnet werden, in dem ein Effekt (mit in der Regel 95-prozentiger Wahrscheinlichkeit) erwartet werden kann, wenn die Ergebnisse einer „Random-Effects“-Metaanalyse auf eine Einzelsituation übertragen werden. Ein Prädiktionsintervall kann auch bei heterogener Studienlage berechnet werden und dann bei der Interpretation der Datenlage hilfreich sein.

Wenn wegen Heterogenität ein Summenschätzer nicht berechnet werden kann, das Prädiktionsintervall den Nulleffekt aber nicht überdeckt, soll nach dem Überarbeitungsentwurf des Methodenpapiers von „deutlich gleichgerichteten Effekten“ ausgegangen werden, die wie signifikante Ergebnisse für den Summenschätzer bei homogener Studienlage behandelt werden. Überdeckt das

Prädiktionsintervall dagegen den Nulleffekt, sollen nur „mäßig gleichgerichtete Effekte“ konstatiert werden, und auch nur dann, wenn weitere Kriterien erfüllt sind. Diese weiteren Kriterien sind weitgehend identisch mit den Kriterien, die bisher für die Feststellung „gleichgerichteter Effekte“ ausreichen (siehe IQWiG Methodenpapier 4.0, Seite 38).

Im Endergebnis werden mit Einführung des Prädiktionsintervalls als Methode höhere Anforderungen gestellt an (dann jedoch auch „deutlich ...“) „gleichgerichtete Effekte“, da Prädiktionsintervalle beispielsweise in der Regel breiter sind als Konfidenzintervalle (2). Die Verwendung von Prädiktionsintervallen kann dazu führen, dass Ergebnisse von Metaanalysen konservativer interpretiert werden (in Richtung eines Nulleffektes), wie auch kürzlich im Rahmen einer empirischen Überprüfung gezeigt worden ist (3). Sie erlauben zudem ein standardisiertes, quantifizierendes und damit objektiveres Vorgehen bei Heterogenität und erscheinen deshalb eine Verbesserung. „Deutlich gleichgerichtete Effekte“ führen gemäß der geplanten Überarbeitung, wie in der Tabelle 2 erkennbar, als Konsequenz zu gleichen Schlussfolgerungen für die Beleglage wie signifikante Ergebnisse für den Summenschätzer bei homogener Studienlage. Nur „mäßig gleichgerichtete Effekte“ würden im Gegensatz dazu – und im Gegensatz zum Verfahren gemäß Methodenpapier 4.0 bei „gleichgerichteten Effekten“ – im Endergebnis zu einer Herabstufung der Beleglage führen: von Beleg zu Hinweis bzw. von Hinweis zu Anhaltspunkt. Auch hier bedeuten die vorgesehenen Änderungen strengere Anforderungen an eine höhere Beleglage.

Da Prädiktionsintervalle erst ab mindestens vier Studien valide berechnet werden können (1), werden bei einem Pool von zwei oder drei Studien gesonderte Kriterien für die Einteilung als „mäßig gleichgerichtete Effekte“ oder „deutlich gleichgerichtete Effekte“ aufgestellt (Seite 16, Punkte 1. und 2.). Diese Kriterien erscheinen plausibel und nachvollziehbar; es ist jedoch unklar, durch welche Überlegungen und empirischen Erfahrungen sie genau begründet sind.

In diesem Punkt unterstützt die AkdÄ den Vorschlag des IQWiG, schlägt jedoch vor, die Kriterien für die Einteilung bei Vorliegen eines Pools von zwei und drei Studien differenzierter darzustellen und mit empirischen Erfahrungen zu unterstützen.

Die gegenüber der entsprechenden Tabelle im Methodenpapier 4.0 jetzt erweiterte Tabelle 2 für die Anforderungen an die Beleglage ergibt sich allein durch die

differenziertere Vorgehensweise für Studien mit „gleichgerichteten Effekten“ bei heterogener Studienlage. Das Verfahren zur Ermittlung der Beleglage ist damit im Überarbeitungsentwurf noch differenzierter und komplexer geworden; inhaltlich erscheint er nachvollziehbar und begründet. Die weitere Differenzierung gerade für Situationen mit heterogenen Studien, die aber gleichgerichtete Effekte aufweisen, ist vor dem Hintergrund der Arzneimittel-Nutzenbewertungsverordnung (AM-NutzenV) notwendig geworden, die als eine der Kategorien einen belegten, aber „nicht quantifizierbaren Zusatznutzen“ vorsieht.
Die AkdÄ stimmt der weiteren Differenzierung zu.

Das Schema der Anforderungen an die Beleglage (Tabelle 2) zielt weiterhin nicht auf die Situationen ab, in denen auf Basis der Ergebnisse einer einzigen Studie Nutzenbelege abgeleitet werden können. Wie im Methodenpapier 4.0 findet sich hierzu weiterhin nur als einziger knapper Satz „Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und ihre Ergebnisse besondere Anforderungen zu stellen [157].“ Die Literaturstelle [157] verweist auf ein Dokument der EMA aus 2011: „Points to consider on application with: 1. meta-analyses; 2. one pivotal study“. Dort werden unter “III.2 Prerequisites for one pivotal study application“ die – explizit besonderen – Anforderungen genannt, die bei Einreichung lediglich einer Studie für eine Zulassung erfüllt sein müssen. Unter den Forderungen sind erwähnenswert: externe Validität und Übertragbarkeit der Ergebnisse; klinische Relevanz der Effektgröße; Signifikanzniveau deutlich unter 5 % mit engen Konfidenzintervallen; innere Konsistenz über Subgruppen; keine Zentreneffekte.
Die Situationen, dass eine einzelne Studie für die Nutzenbewertung eines Arzneimittels zur Verfügung steht, sind, wie die bisherigen Verfahren gezeigt haben, nicht selten (z. B. Ticagrelor, Apixaban bei VHF, demnächst Rivaroxaban und Dabigatran bei VHF, viele onkologische Wirkstoffe). Es ist nicht nachvollziehbar, warum entsprechende Anforderungen bisher nicht weiter konkretisiert und operationalisiert worden sind und dies bei der Überarbeitung ebenfalls nicht vorgesehen ist. Zudem beziehen sich die Anforderungen der EMA in dem Dokument auf die Zulassung eines Wirkstoffs und es ist unklar, ob und wie sie für die Nutzenbewertung anwendbar sind.

Für Situationen, in denen für die Nutzenbewertung nur eine einzige Studie vorliegt, schlägt die AkdÄ vor, eine von den Kriterien der Zulassung unabhängige Operationalisierung zu erarbeiten.

Ad 3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V
(**einschließlich Anhang: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens**)

Im Zentrum dieses Abschnittes einschließlich des entsprechenden Anhangs steht die Operationalisierung zur Feststellung des Ausmaßes des Zusatznutzens. Ziel dieser Operationalisierung ist die Umsetzung von Rechtsvorgaben, die in § 5 Abs. 7 der AM-NutzenV formuliert sind. Diese Verordnung gibt sowohl für das Ausmaß (erheblich; beträchtlich; gering; nicht quantifizierbar; kein; geringerer) als auch für Qualität und Art des Zusatznutzens (z. B Überlebensdauer; schwerwiegende Symptome; Lebensqualität) Kategorien vor, auch wenn diese nicht abschließend formuliert oder konkretisiert sind. Die Aktualisierung des Methodenpapiers ist auf der einen Seite bei der Operationalisierung des Ausmaßes des Zusatznutzens an diese Vorgaben gebunden, muss sie andererseits für die Bewertung im konkreten Fall umsetzbar machen.
Weiter ist von Bedeutung, dass der Gesetzgeber für die Preisfindung zur Erstattung von (zumindest neu zugelassenen) Arzneimitteln im GKV-Bereich primär nicht Kosten-Nutzen- oder Kosten-Nutzwert-Analysen vorsieht, sondern eine vorgeschaltete Nutzenanalyse mit nachfolgender Festsetzung bzw. Verhandlung des Erstattungsbetrags. Dieses Verfahren ist zunächst zu begrüßen; für seine Umsetzung existieren aber international vergleichsweise nur wenige verwertbare Erfahrungen.

Die im Überarbeitungsentwurf dargestellte Operationalisierung zur Feststellung des Ausmaßes des Zusatznutzens ist *de facto* nicht neu; die Kernelemente sind bereits in der Dossierbewertung von Ticagrelor durch das IQWiG beschrieben (Anhang A Dossierbewertung A11-02 vom 29.09.2011). Dieses Verfahren zur Operationalisierung wendet das IQWiG seitdem bei Nutzenbewertungen von Arzneimitteln nach § 35a SGB V an.

Von den in der AM-NutzenV vorgegebenen Kategorien für das Ausmaß des Zusatznutzens sind für die Operationalisierung nur die Kategorien „erheblich“, „beträchtlich“ und „gering“ von Bedeutung. Nur patientenrelevante Nutzen- und Schadenereignisse (wenn man von den seltenen Ausnahmen solcher Surrogate absieht, die ausreichend validiert sind und gleichzeitig eine Quantifizierung patientenrelevanter Ereignisse zulassen) werden betrachtet und hierarchisiert. Die Ausgestaltung der Kategorien und Einordnung einzelner Nutzen- oder Schadenereignisse in die Kategorien ist im Detail arbiträr, grundsätzlich aber nachvollziehbar. Weil die Einteilung im Detail arbiträr ist, erscheint wichtig, dass sie transparent erfolgt, um gegebenenfalls andere Einteilungen und Bewertungen von Nutzen- oder Schadenereignissen zu ermöglichen. Letzteres erscheint in ausreichendem Maße gegeben.

Weiterhin werden für jede der Ereigniskategorien spezifische Grenzen zum Ausmaß von Effekten festgelegt, die eine Kategorisierung als „erheblicher“, „beträchtlicher“ oder „geringer“ Zusatznutzen erlauben. Die Kategorisierung des Ausmaßes des Zusatznutzens werden exemplarisch für relative Risiken und Hazard Ratios entwickelt (und versucht, andere Effektmaße auf diese zurückzuführen), wobei für die Patienten positive Veränderungen von Ereignisraten als relative Risiken oder Hazard Ratios unter 1 dargestellt werden. Die relativen Risiken (genauer die oberen Grenzen des 95 % Konfidenzintervalls), die ein Arzneimittel für einen beispielsweise beträchtlichen Zusatznutzen erreichen muss, sind bei Ereignissen, die für die Patienten gravierend sind, näher an der 1 gelegen (Nulleffekt) als bei nicht schwerwiegenden Ereignissen. Auch die Grenzsetzung in den einzelnen Ereigniskategorien für einen „erheblichen“, „beträchtlichen“ oder „geringen“ Zusatznutzen ist arbiträr; andere Grenzen könnten diskutiert werden. Soweit bekannt, liegen empirische Erfahrungen mit dieser Vorgehensweise auch aus anderen Ländern kaum vor; dies erscheint plausibel und begründet.

Es wird in dem Methodenpapier weiterhin von der Grundforderung ausgegangen, dass hinsichtlich der Kategorie Mortalität eine Halbierung des Risikos gegenüber einer Vergleichstherapie erreicht werden muss, um als erheblicher Zusatznutzen bezeichnet werden zu können (4). Dabei handelt es sich aber um eine arbiträre Grenzsetzung, die nicht weiter auf Konsistenz und Anwendbarkeit geprüft ist. Aus dieser Grundannahme entwickelt das IQWiG mittels verschiedener Rasterungen für die einzelnen Kategorisierungen Grenzen (Schwellenwerte) für die übrigen

Kategorien zum Ausmaß des Zusatznutzens bei der Zielgröße Überleben und auch die Grenzen für das Ausmaß „erheblicher“, „beträchtlicher“ oder „geringer“ Zusatznutzen für andere Kategorien der Zielgrößen. In Simulationsverfahren wurde dann gezeigt, dass die in dem Überarbeitungsentwurf verwendeten Schwellenwerte zu Kategorien für das Ausmaß des Zusatznutzens führen, die (in praktisch relevanten Bereichen) unabhängig vom Basisrisiko der untersuchten Populationen sind.

Die verwendeten Grenzziehungen selbst – von den dann zur Ausgestaltung verwendeten Methoden abgesehen – sind kein methodisches Problem, sondern die Wertungen, die in dem Geltungsraum von Entscheidungen, die daraus abgeleitet werden, konsentiert sein müssen. Wenn sie auch grundsätzlich plausibel erscheinen, sind jetzt allerdings schon dadurch Tatsachen geschaffen worden, dass sie – zumindest vom IQWiG für seine Dossierbewertungen – bei der frühen Nutzenbewertung Anwendung gefunden haben. Für zukünftige Bewertungen werden diese Grenzziehungen aufgrund des Gebots der Gleichbehandlung schwer rückgängig zu machen sein.

Für die Formulierung des Ausmaßes des Zusatznutzens wird auf das relative Risiko als Effektmaß zurückgegriffen und nicht auf absolute Risikodifferenzen. Die Vorteile des relativen Risikos gegenüber Risikodifferenzen werden im Entwurf der Überarbeitung ausführlich beschrieben (beispielsweise Stabilität des relativen Risikos gegenüber dem Basisrisiko; Abhängigkeit der Risikodifferenzen von betrachteten Zeitintervallen). Wenn es gilt, Methoden (hier speziell Wirkstoffe) zu bewerten, ist dies unstrittig. Allerdings sollte auch bewertet oder zumindest ausreichend transparent dargestellt werden, welche absoluten Ausmaße an Zusatznutzen sich hinter den einzelnen Kategorien „erheblicher“, „beträchtlicher“ oder „geringer“ Zusatznutzen für die deutsche Versorgungssituation verbergen. Als Beispiel würde eine 50-prozentige Reduktion der Mortalität durch ein Arzneimittel (erheblicher Zusatznutzen) genauso eine Reduktion von 10 % auf 5 % wie von 1 % auf 0,5 % bedeuten können. Die entsprechenden Risikodifferenzen sind aber – trotz gleicher Kategorisierung des Zusatznutzens nach der Methode des IQWiG – unter Versorgungsaspekten unterschiedlich zu bewerten. Der Vorteil der relativen Risiken wie beispielsweise Übertragbarkeit auf ganze oder Subpopulationen mit anderem Basisrisiko soll nicht negiert werden.

Die AkdÄ regt an, zu überlegen, ob bei den endpunktbezogenen Ergebnissen abschließend neben den Ausmaßkategorien („erheblicher“, „beträchtlicher“ oder

„geringer“ Zusatznutzen) auch die relativen Risiken (mit 95 % Konfidenzintervall) und die absoluten Risikodifferenzen (mit 95 % Konfidenzintervall) für die untersuchte Population und Zeitdauer angegeben werden sollten. Dies würde die Transparenz maßgeblich erhöhen, ohne eine Übertragung auf Sub- oder andere Populationen zu verhindern.

Des Weiteren schlägt die AkdÄ vor, zu erwägen, ob bei der Ergebnisdarstellung für den Zusatznutzen bei Arzneimitteln auf eine abschließende zusammenfassende Bewertung nicht verzichtet werden kann und nur das Ausmaß des Zusatznutzens für die einzelnen Endpunkte präsentiert werden sollte. Die abschließende Bewertung unter Abwägung der Nutzen- und Schadensaspekte ist (zumindest bisher) kein methodisches Verfahren, sondern eine Werteabwägung, für die ein Konsensprozess unter den Betroffenen notwendig ist.

Die AkdÄ regt an, eventuell auf die abschließende zusammenfassende Bewertung zu verzichten, zumindest aber die endpunktbezogenen Ergebnisse, wie oben beschrieben, transparenter und mit den absoluten Risikodifferenzen (einschließlich der bezüglichen Rahmenbedingungen) darzustellen.

Ad 7.3.8 Meta-Analysen

Der Überarbeitungsentwurf unterscheidet sich vom entsprechenden Abschnitt in dem Methodenpapier 4.0 lediglich dadurch, dass 2 Absätze ergänzt werden sollen, die sich mit der geplanten Verwendung der Prädiktionsintervalle befassen. Demnach ist vorgesehen, Prädiktionsintervalle grafisch künftig als zusätzliche waagerechte Rechtecke in Forest-Plots aufzunehmen, um die Heterogenität der Studien zu veranschaulichen. Die Prädiktionsintervalle sollen dabei nicht zur Beurteilung der Signifikanz von Effekten herangezogen werden.

Die AkdÄ stimmt dieser Ergänzung zu.

Literatur

1. Higgins JP, Thompson SG, Spiegelhalter DJ: A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc 2009; 172: 137-159.

2. Riley RD, Higgins JP, Deeks JJ: Interpretation of random effects meta-analyses. BMJ 2011; 342: d549.

3. Graham PL, Moran JL: Robust meta-analytic conclusions mandate the provision of prediction intervals in meta-analysis summaries. J Clin Epidemiol 2012; 65: 503-510.

4. Djulbegovic B, Kumar A, Soares HP et al.: Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006. Arch Intern Med 2008; 168: 632-642.

A.1.3 – Bayer Vital GmbH

Bayer HealthCare

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Im Mediapark 8
50670 Köln

Stellungnahme der Bayer Vital GmbH zur Aktualisierung der Allgemeinen Methoden des IQWiG

21. Mai 2013

Michael Meinhardt

Bayer Vital GmbH
Market Access
Health Economics &
Outcomes Research
Gebäude K 56
51366 Leverkusen
Deutschland

[redacted]
www.bayervital.de

Geschäftsführer:
Frank Schöning

Vorsitzender des
Aufsichtsrats:
Manfred Vehreschild

Sitz der Gesellschaft:
Leverkusen
Amtsgericht Köln
HRB 49226

Diese Stellungnahme bezieht sich auf die „Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden 4.1“ des IQWiG (Entwurf vom 18.04.2013) und hier konkret auf folgende aktualisierte bzw. neue Abschnitte:

- 3.1.4 Endpunktbezogene Bewertung,
- 3.1.5 Zusammenfassende Bewertung,
- 3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V,
- Neuer Anhang: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens.

Vorwort

Im Rahmen der frühen Nutzenbewertung nach § 35a SGB V stand das IQWiG als bewertendes Institut vor der Herausforderung die Bewertung des Ausmaßes und der Wahrscheinlichkeit eines Zusatznutzens von Arzneimitteln zu operationalisieren. Entsprechend der gesetzlichen Vorgaben kann das Ausmaß des Zusatznutzens drei Ausprägungen aufweisen: erheblich, beträchtlich oder gering. Diese kategorial ordinale und verbale Einteilung ist in der Arzneimittel-Nutzenbewertungsverordnung (AM-NutzenV) vorgegeben. Das IQWiG hat einen konzeptionellen Vorschlag zur Klassifizierung und Quantifizierung des Zusatznutzens sowie zur Beurteilung der Aussagesicherheit entwickelt. BAYER begrüßt den Vorschlag des IQWiG grundsätzlich und respektiert die Pionierarbeit, die das IQWiG in diesem Bereich leistet. Gleichzeitig möchte BAYER auf einige aus seiner Sicht problematische Aspekte hinweisen und das IQWiG um erläuternde Ergänzungen bitten.

Hierarchisierung von Zielgrößen

Zielgrößen werden vom IQWiG in einer Rangliste in Form von Zielgrößenkategorien aufgeführt, wobei die Mortalität grundsätzlich an erster Stelle steht (Kategorie 1), gefolgt von schwerwiegenden Symptomen und Nebenwirkungen sowie der Lebensqualität (Kategorie 2) und schließlich nicht schwerwiegenden Symptomen bzw. Nebenwirkungen (Kategorie 3). Die in der AM-NutzenV genannten Kriterien für die Zuteilung in einen erheblichen, beträchtlichen und geringen Zusatznutzen sehen die Zuteilung in dieser Form nicht vor.

Das IQWiG begründet sein Vorgehen wie folgt: *„Die in […] der AM-NutzenV gelieferten Kriterien für das Ausmaß des Zusatznutzens benennen (Rechts-) Begriffe, die zum Teil eindeutig bestimmt […], teilweise weniger eindeutig bestimmt sind […]. Darüber hinaus sind die Kategorien nicht für alle aufgeführten Kriterien erschöpfend besetzt, z. B. werden für die „Überlebensdauer“ nur Beispiele für die Kategorien „erheblicher“ und „beträchtlicher“ Zusatznutzen genannt. […] In einem ersten Schritt ist es also sinnvoll, die Kriterienliste anzupassen und durch qualitativ und quantitativ gleichwertige Kriterien zu ergänzen. […] Ausgehend von diesen Ergänzungen ist eine Umstrukturierung der Zielgrößenkategorien angezeigt, um die in der AM-NutzenV intendierte Hierarchisierung der Zielgrößen abzubilden und gemäß § 5 Abs. 7 der AM-NutzenV den Schweregrad der Erkrankung zu berücksichtigen“.*

Das Vorgehen des IQWiG ist prinzipiell nachvollziehbar und grundsätzlich ist es zu begrüßen, wenn Endpunkte entsprechend ihrer Bedeutung hierarchisiert werden, so dass der Schweregrad der Erkrankung Berücksichtigung findet. Allerdings ist genau dieser Aspekt im Operationalisierungsvorschlag des IQWiG aus Sicht von BAYER nicht adäquat abgebildet. Da die Bedeutung verschiedener Endpunkte ja gerade abhängig von der Art und dem Schweregrad der Erkrankung ist, läuft eine solche generalisierte Hierarchisierung von Zielgrößen dieser Idee zuwider. Die grundsätzliche Priorisierung des Endpunktes Mortalität wird als problematisch erachtet, da der Erhalt der Lebensqualität, beispielsweise in palliativen Therapiesituationen im Rahmen von Krebserkrankungen, eine herausragende Rolle spielen kann. So wird laut der Deutschen Gesellschaft für Hämatologie und Onkologie (DGHO) und der deutschen Krebsgesellschaft (DKG) einer relativ kurzen Lebenszeitverlängerung von Betroffenen in manchen Erkrankungssituationen nicht unbedingt mehr Wert beigemessen als dem Erhalt der Lebensqualität oder der Vermeidung schwerwiegender Nebenwirkungen (1). Auch in einem Indikationsgebiet, in dem es z. B. um die Verhinderung von Erblindungen geht (ohne Lebensbedrohung), ist es ebenfalls schwer nachvollziehbar, warum die Zielgrößenkategorie 1 dann entsprechend gar nicht besetzt sein soll. Eine Erblindung stellt für die Betroffenen ein ausgesprochen relevantes Ereignis dar, das es in höchster Priorisierung zu verhindern gilt. Zum Erreichen eines erheblichen Zusatznutzens müssen in dem Fall aber sofort die erschwerten Hürden in Form kleinerer zu unterschreitender Schwellenwerte erreicht werden, da Erblindung „nur“ in die Zielgrößenkategorie 2 fallen würde.

Der bisherige Vorschlag stellt ein „One Size fits all“-Konzept dar, welches aus Sicht von BAYER den komplexen Anforderungen in der Bewertung des Zusatznutzens von Arzneimitteln nicht ausreichend gerecht wird. Die Methodik des IQWiG und die im Rahmen der Operationalisierung geforderten Effektstärken berücksichtigen weder verschiedene Erkrankungsbilder oder Krankheitsverläufe noch verschiedene Krankheitsschweren oder Erkrankungsstadien. BAYER stellt daher in Frage, ob eine solche indikationsübergreifende Operationalisierung sinnvoll zur Bewertung von allen Arzneimitteln eingesetzt werden kann. In jedem Falle sollte die vorgenommene generalisierte Hierarchisierung von Zielgrößen vom IQWiG zumindest begründet werden.

Wahl von Effektmaßen und Effektstärken

Das IQWiG schreibt: *„Zu den Fragen, welche Effektstärken für die einzelnen Zielgrößen zu welcher Ausmaßkategorie führen und welche Effektmaße für diese Bewertung zu wählen sind, finden sich in der AM-NutzenV keine Angaben. Diese Fragen können prinzipiell nur bedingt methodisch beantwortet werden. Dennoch besteht die Notwendigkeit, das in den Dossiers dargelegte Ausmaß des Zusatznutzens zu bewerten […] und selbst Aussagen zum Ausmaß zu machen. Um hierbei die im weiteren Abwägungsprozess notwendigerweise zu treffenden Werturteile möglichst gering zu halten und diese explizit zu machen, bedarf es einer*

- *expliziten Operationalisierung, um ein transparentes und nachvollziehbares Verfahren sicherzustellen, sowie einer*
- *abstrakten Operationalisierung, um größtmögliche Konsistenz zwischen den Nutzenbewertungen zu erzielen.“*

Der Operationalisierungsvorschlag des IQWiG ermöglicht insofern ein sehr transparentes Nutzenbewertungsverfahren, als dass das Ergebnis zu dem das IQWiG im Rahmen seiner Bewertung kommt, durch die präsentierte Bewertungsmatrix sehr vorhersehbar wird. BAYER ist grundsätzlich an einem pragmatischen Bewertungsverfahren interessiert, doch wird dieser Ansatz der Komplexität der zu bewältigenden Aufgabe nicht gerecht. Auch genügt es dem Anspruch einer größtmöglichen Konsistenz zwischen den Nutzenbewertungen nicht, wenn ein starres Bewertungsschema über alle Indikationen und individuelle krankheitsspezifische Besonderheiten hinweg angewendet wird. BAYER ist daher der Meinung, dass die Bewertung eines Zusatznutzens nicht ausschließlich auf Basis eines einzelnen, aus dem Zusammenhang gerissenen, statistischen Ergebnisses erfolgen sollte. Zudem beinhaltet die Bewertungsmatrix des IQWiG auch keinen Vorschlag für den Umgang mit nicht-binären Zielkriterien, bei denen die jetzige Methode nicht direkt angewendet werden kann. Bei aller „Quantifizierungslust“ muss daher ein gesundes und individuelles Abwägen erfolgen, ob der entwickelte Operationalisierungsvorschlag uneingeschränkt Anwendung finden kann.

Das IQWiG verweist zudem selber darauf, dass die Operationalisierung des Zusatznutzens notwendigerweise an das Treffen von Werturteilen gebunden ist. An einen Operationalisierungsvorschlag sind notwendigerweise statistische Setzungen gekoppelt, und es ist bekannt, dass solche auf den ersten Blick statistischen Setzungen (wie z. B. ein p-Wert von <0,05 zum Nachweis von Signifikanz) unvermeidlich mit Werturteilen verbunden sind. Diese Tatsache allein ist aus Sicht von BAYER nicht kritikfähig, da Werturteile bei Konsens und aus pragmatischen Gründen durchaus angemessen sein können. Im Rahmen der Entwicklung des Operationalisierungsvorschlags fehlte aber eine Beteiligung und Diskussion innerhalb der Fachöffentlichkeit und mit Betroffenen, sodass letztlich die Legitimation zur Entscheidung bezüglich wichtiger normativer Fragen fehlt.

Verwendung von relativen Risiken als Effektmaß

Das IQWiG sah sich vor der Herausforderung ein geeignetes Effektmaß im Rahmen der Operationalisierung des Zusatznutzens zu wählen. BAYER begrüßt, dass das IQWiG auf seine Beweggründe relative Effektmaße zu verwenden, in der Aktualisierung des Methodenpapiers näher eingeht als dies noch im Anhang A zur Nutzenbewertung von Ticagrelor der Fall war (2).

Bei der Wahl fokussierte sich das IQWiG nach eigenen Angaben zunächst auf die Situation binärer Daten. Dies ist zwar nachvollziehbar, allerdings wirft es sogleich die Frage auf, ob die Operationalisierung in Situationen mit nicht-binären Daten in gleicher Form Anwendung finden kann.

Als Gründe zur Wahl von relativen Effektmaßen nennt das IQWiG deren Vorteile gegenüber absoluten Effektmaßen. Diesen Vorteilen kann seitens von BAYER grundsätzlich gefolgt werden, allerdings weist die Verwendung relativer Effektmaße gleichzeitig auch erhebliche Nachteile auf, die vom IQWiG weder genannt noch diskutiert werden.

Bereits im Rahmen des Stellungnahmeverfahrens zur Nutzenbewertung von Ticagrelor wurde die alleinige Verwendung eines relativen Effektmaßes von vielen Seiten kritisiert. So wurde beispielsweise von der Arzneimittelkommission der deutschen Ärzteschaft (AkdÄ) angemerkt, dass es intuitiv schwer nachvollziehbar sei, warum die Reduktion einer Mortalitätsrate von 10% auf 5% als „erheblich“, eine Reduktion von 35% auf 30% aber nur als „gering“ eingestuft werde (3). Mit dem Quantifizierungsvorschlag des IQWiG sei keine vergleichende Bewertung möglich. Je gravierender eine Ausgangssituation ist (also je höher ein bestehendes Risiko), umso höher muss die Risikoreduktion sein, um einen erheblichen Zusatznutzen attestiert zu bekommen. Relative Risiken allein können die klinische Relevanz nicht immer adäquat widerspiegeln. Diese Kritik entkräftet das IQWiG in seinen aktualisierten Methoden nicht.

Da offensichtlich kein Effektmaß allein geeignet ist, die klinische Relevanz einer Therapieverbesserung widerzuspiegeln, sollte die Verwendung verschiedener Effektmaße diskutiert werden. Röhmel (2012) schlägt z. B. die „Number Needed to Treat“ als Effektmaß vor, um den Unterschied basierend auf absoluten Unterschieden darzustellen (4). Zu dieser Forderung äußert sich das IQWiG ebenfalls nicht.

Es argumentiert lediglich wie folgt: *„Zusammenfassend mögen absolute Risikoreduktionen in einer individuellen Entscheidungssituation eher handlungsleitend sein, für allgemeine Aussagen im Sinne einer Bewertung des Zusatznutzens eines Arzneimittels sind dagegen relative Effektmaße besser geeignet.“*

Dieser Argumentation kann BAYER nicht folgen, da es sich im Falle der Bewertung von Arzneimitteln in der Tat sehr häufig um ganz individuelle Entscheidungssituationen handelt und die Wahl geeigneter Effektmaße daher keine „Entweder-Oder-Entscheidung“ sein sollte. Einer allein auf relativen Risiken basierenden Operationalisierung fehlt in jedem Fall die notwendige Flexibilität, um verschiedenen Indikationsgebieten und vor allem der Bedeutung von absoluten Unterschieden gerecht zu werden (4).

Verwendung eines Streuungsmaßes als Effektmaß

Zudem erachtet BAYER es als kritisch, dass die Operationalisierung des Zusatznutzens ausschließlich durch ein Streuungsmaß erfolgt.

Das IQWiG erläutert hierzu: *„Das Konzept sieht vor, dass ein 95%-Konfidenzintervall im Sinne einer verschobenen Hypothesengrenze einen jeweiligen Schwellenwert unterschreitet, damit das Ergebnis als erheblicher, beträchtlicher oder geringer Zusatznutzen eingestuft wird. Ein solches inferenzstatistisches Vorgehen hat gegenüber der Betrachtung von Punktschätzern zwei wesentliche Vorteile: (i) Die Präzision der Schätzung fließt in die Bewertung ein; (ii) Die statistischen Irrtumsmöglichkeiten lassen sich damit einhergehend auf übliche kleine Werte (z. B. 5%) beschränken.“*

Die Begrenzung der Irrtumswahrscheinlichkeit bei der Quantifizierung des Zusatznutzens ist ein nachvollziehbares Anliegen. Die obere Grenze des Konfidenzintervalls als Basis zur Einordnung in eine Nutzenausmaßkategorie zu verwenden, kann aber dazu führen, dass eigentlich schwächere Effekte, die mit einer größeren Fallzahl gezeigt wurden, einen höheren Nutzen attestiert bekommen, als stärkere Effekte, die mit kleiner Fallzahl gezeigt wurden. Eine Bewertung des Zusatznutzens anhand der Konfidenzintervallgrenze bedeutet damit, dass der eigentliche Effektschätzer fallzahlabhängig gemacht wird. Seltenere Erkrankungen bei denen Studienpopulationen für gewöhnlich kleiner sind, werden dadurch systematisch benachteiligt, da der eigentlich beobachtete Effekt außer Acht gelassen wird. BAYER ist daher der Meinung,

dass im Sinne der Patientenrelevanz auch der Punktschätzer mit in die Bewertung des Zusatznutzens einbezogen werden sollte.

Darüber hinaus bedeutet die Verwendung eines Streuungsmaßes als Effektmaß, dass das Ausmaß des Zusatznutzens zusätzlich abhängig von der Aussagesicherheit gemacht wird. Die Aussagesicherheit soll ja aber durch die Wahrscheinlichkeit des Zusatznutzens anhand der Begriffe „Beleg“, „Hinweis“ und „Anhaltspunkt“ abgebildet werden.

Festlegung der geforderten Effektstärken

Das IQWiG legt der Berechnung der Schwellenwerte für die obere Grenze der Konfidenzintervalle gewünschte Effektgrößen zu Grunde. Die gesamte Bewertungsmatrix basiert letztlich nur auf einer Literaturquelle – einer Veröffentlichung aus dem Jahr 2008, in der onkologische Therapiestudien ausgewertet und schließlich diejenigen, bei denen das relative Risiko des Endpunkts Mortalität 0,5 oder weniger betrug als „breakthrough interventions“ bezeichnet wurden (5). Bei dieser einzigen expliziten Referenz handelt es sich um das stellvertretende Urteil einer Autorengruppe, die die Grenze von 0,5 selbst als „arbiträr gewählt“ bezeichnet. Dennoch nimmt das IQWiG diese Arbeit zum Anlass, für einen „erheblichen Zusatznutzen“ in der Kategorie „Mortalität“ als Effektstärke grundsätzlich ein relatives Risiko von 0,5 oder weniger vorauszusetzen.

Die Herleitung aller übrigen Effektstärken erfolgt schließlich sogar ohne jegliche wissenschaftliche Basis. Während im Anhang A zur Nutzenbewertung von Ticagrelor seitens des IQWiG gar keine Begründung für das Vorgehen geliefert wurde, so lautet die Erläuterung in der Aktualisierung der Methoden hierzu: *„Im nächsten Schritt mussten dann für die Ausmaßmatrix die übrigen tatsächlichen Effekte festgelegt und die dazugehörigen Schwellenwerte ermittelt werden. Dabei war zu beachten, dass die Anforderungen von der Zielgrößenkategorie „Mortalität“ ausgehend für weniger schwerwiegende Zielgrößen zunehmen und von der Ausmaßkategorie „erheblich“ ausgehend für niedrigere Ausmaßkategorien abnehmen sollten. Eine Rasterung von 1/6 für die tatsächlichen Effekte erwies sich dabei als pragmatische Lösung.“*

Diese Herangehensweise ist unter Anbetracht des Zeitdrucks und des akuten Regelungsbedarfs im Rahmen der ersten Nutzenbewertung, die das IQWiG vorzunehmen hatte, zwar verständlich, zugleich scheint die Festlegung aber auch willkürlich. Im Rahmen des Stellungnahmeverfahrens zur ersten Nutzenbewertung von Ticagrelor gingen bereits zahlreiche kritische Stellungnahmen dazu ein. So wurde z. B. von der DGHO und der DKG umfangreich kommentiert, dass der Operationalisierungsvorschlag des IQWiG intransparent und nicht ausreichend wissenschaftlich begründet sei (1). Die AkdÄ und der Verband Forschender Arzneimittelhersteller (VFA) führten an, dass die geforderten Effekte arbiträr

gewählt und nicht aus einer abgeschlossenen wissenschaftlichen Diskussion heraus entstanden seien und die Setzungen und Werturteile nicht den Standards der evidenzbasierten Medizin entsprächen, sondern nicht nachvollziehbar hergeleitete Eigenkonstrukte seien (3).

Die scheinbare Willkürlichkeit in der Festlegung von Effektstärken ohne vorhergehende öffentliche Diskussion wird von BAYER als besonders problematisch empfunden. So wird das vom IQWiG geforderte relative Risiko von 0,5 für den Endpunkt Mortalität von der AkdÄ gar als „realitätsfern“ bezeichnet (3). Zur grundsätzlich schweren Erreichbarkeit einer sehr großen therapeutischen Verbesserung kommt die Schwierigkeit, einen solchen Effekt auch zu zeigen. In den Zulassungsstudien, welche zum frühen Zeitpunkt der Bewertung vorliegen, geht es in erster Linie um die Arzneimittelsicherheit und -wirksamkeit und nicht um die Größe eines Effekts. Gerade wenn sich die eindeutige Überlegenheit einer neuen Substanz schon in einer Interimsanalyse zeigt, wird die entsprechende Studie in der Regel frühzeitig abgebrochen. Die Fortführung wäre z. B. bei onkologischen Studien nicht vertretbar, geschweige denn die Wiederholung einer solchen Studie zur Verbesserung der Beleglage. Es darf grundsätzlich auch nicht verlangt werden, dass Studien größer sind als gerade nötig, um die klinische Relevanz und die statistische Signifikanz übereinzubringen, und mehr Patienten unter Risiko zu stellen als notwendig, um damit Erkenntnisse über die Größe des Effektes zu erhalten (6).

Dies würde bedeuten, dass bei bereits nachgewiesener Überlegenheit beispielsweise ein Cross-Over der Patienten vom Kontrollarm in den aktiven Arm unterbunden und Patienten der schlechteren Therapie ausgesetzt werden müssten, um die Aussagesicherheit und die Effektstärke zu erhöhen. Eine gravierende Verbesserung wie die Halbierung der Mortalität würde in der entsprechenden Zulassungsstudie möglicherweise gar nicht gezeigt werden können (zumindest nicht mit engem Konfidenzintervall), da die Fallzahlen bei vorzeitiger Beendigung entsprechend wesentlich kleiner ausfielen und die Power der Studie geringer wäre.

BAYER fordert daher, dass grundsätzlich diskutiert wird, ob es nicht im Widerspruch zum Gedanken der frühen Nutzenbewertung steht, wenn zu einem solch frühen Zeitpunkt der Zulassung eines Wirkstoffs, bereits eine qualitativ und quantitativ sehr hochwertige Studienlage erwartet wird.

Schaden-Nutzen-Abwägung/ Ableitung einer Gesamtaussage

Das IQWiG steht vor der Herausforderung, die verschiedenen Nutzenausmaße in einzelnen Endpunkten zu einer Gesamtwertung zu aggregieren und im Rahmen dessen auch Nutzen und Schaden gegeneinander abzuwägen.

Das IQWiG führt hierzu in Abschnitt 3.1.5 aus: *„Eine Möglichkeit der gleichzeitigen Würdigung von Nutzen und Schaden ist die Gegenüberstellung der endpunktbezogenen Nutzen- und Schadenaspekte. Dabei werden die Effekte auf alle Endpunkte (qualitativ oder semiquantitativ) gegeneinander abgewogen mit dem Ziel, zu einer endpunktübergreifenden Aussage zum Nutzen bzw. Zusatznutzen einer Intervention zu kommen. Eine weitere Möglichkeit der gleichzeitigen Würdigung besteht darin, die verschiedenen patientenrelevanten Endpunkte zu einem einzigen Maß zu aggregieren. […]“*

Am derzeit noch gültigen Methodenpapier des IQWiG (4.0) wurde seinerzeit scharf kritisiert, dass darin kein Algorithmus beschrieben wird, wie ein Summenparameter berechnet werden kann, der eine Gewichtung von Nutzen und Schaden vornimmt. Das IQWiG hat daraufhin zwei Generalaufträge in Form von Pilotstudien in Auftrag gegeben, mit denen Verfahren entwickelt werden sollten, die es ermöglichen einen „kardinalen Nutzenwert zu beschreiben, mit dessen Hilfe alle denkbaren Therapiealternativen in einem Indikationsgebiet über alle relevanten Endpunkte hinweg verglichen werden können“ (7). Diese beiden Verfahren (Analytic Hierarchy Process bzw. Conjoint-Analyse) werden in der Aktualisierung der Allgemeinen Methoden erwähnt und kurz vorgestellt.

Aus Sicht von BAYER erscheinen die vorgestellten Methoden zur Hierarchisierung und Abwägung verschiedener Endpunkte grundsätzlich sinnvoll. Fraglich bleibt, warum eine solche Hierarchisierung erst im Rahmen der Abwägung zum Gesamtnutzen zum Einsatz kommen soll und nicht bereits die Basis einer indikationsabhängigen Hierarchisierung von Zielgrößen und den entsprechend geforderten Effektgrößen bildet. Wie bereits angemerkt, wäre es wünschenswert, wenn die Zielgrößenkategorien indikationsabhängig entsprechend der tatsächlichen Bedeutung für die Patienten hierarchisiert würden.

Eine explizite Beschreibung des Vorgehens und der Anwendung der beschriebenen Methoden im Rahmen der frühen Nutzenbewertung erfolgt im Methodenpapier nicht. Auch bleibt unklar, wann welches der beiden Verfahren Anwendung finden soll. Aufgrund der unterschiedlichen methodischen Herangehensweisen können die beiden Verfahren aber durchaus voneinander abweichende Ergebnisse liefern.

Das methodische Vorgehen zur Schaden-Nutzen-Abwägung und zur Ableitung eines Gesamtnutzens im Rahmen der frühen Nutzenbewertung wird laut IQWiG weiterhin ohne zugrunde liegende Systematik vorgenommen. In Abschnitt 3.3.3 beschreibt das IQWiG sein Vorgehen wie folgt: *„Für den dritten Schritt der Operationalisierung, der Gesamtaussage zum Ausmaß des Zusatznutzens bei gemeinsamer Betrachtung aller Endpunkte, ist eine strenge Formalisierung nicht möglich, da für die hierzu zu treffenden Werturteile gegenwärtig keine Abstraktion bekannt ist. Das Institut wird im Rahmen seiner Nutzenbewertung die Aussagen zur Wahrscheinlichkeit und zum Ausmaß der Effekte vergleichend gegenüberstellen und einen begründeten Vorschlag für eine Gesamtaussage unterbreiten.“*

Das Vorgehen des IQWiG im Rahmen bisheriger Nutzenbewertungsverfahren war teilweise zwar intuitiv nachvollziehbar, in den Stellungnahmeverfahren einiger Nutzenbewertungen (z. B. Cabazitaxel) wurde es aber auch als nicht nachvollziehbar bezeichnet, wie einzelne Komponenten der Bewertung gewichtet wurden, um zu einer Gesamtaussage zu kommen (8). So wird das Vorgehen des IQWiG z. B. laut DGHO dem Anspruch einer methodisch sauberen und wissenschaftlich fundierten Nutzenbewertung nicht gerecht (8). Die Argumentation, dass für die hierzu zu treffenden Werturteile keine Abstraktion bekannt sei, ist unter Anbetracht der Tatsache, dass der Vorschlag des IQWiG zur Operationalisierung des Zusatznutzens von zahlreichen Werturteilen geprägt ist, überraschend.

Es ist darüber hinaus auch nach wie vor nicht festgelegt wie die ordinal skalierten Begriffe „Beleg“, „Hinweis“ und „Anhaltspunkt“ qualitativ zu bewerten sind. Die „Rangordnung“ in der Mischung aus Ausmaß und Wahrscheinlichkeit des Zusatznutzens ist weiterhin nicht geklärt.

Bemühungen in diesem Feld wären sehr begrüßenswert, da ansonsten eine aufwändig erarbeitete Operationalisierung, welche Nutzen und Schaden je Endpunkt darstellt, durch eine unsystematische Aggregation „neutralisiert“ wird.

Beleglage/ Aussagesicherheit

Neben einer Einstufung des Ausmaßes des Zusatznutzens macht das IQWiG auch Angaben zur Wahrscheinlichkeit des Zusatznutzens und trifft damit eine Aussage zur Ergebnissicherheit. Hierzu führt das IQWiG im neuen Abschnitt 3.1.4 aus: *„Ein wichtiges Kriterium zur Ableitung von Aussagen zur Beleglage ist die Ergebnissicherheit. Grundsätzlich ist jedes Ergebnis einer empirischen Studie oder einer systematischen Übersicht über empirische Studien mit Unsicherheit behaftet und daher auf seine Ergebnissicherheit zu prüfen. Hierbei ist zu unterscheiden zwischen qualitativer und quantitativer Ergebnissicherheit. […] Bei der Ableitung der Beleglage für einen Endpunkt sind die Anzahl der vorhanden Studien, deren qualitative Ergebnissicherheiten sowie die in den Studien gefundenen Effekte von zentraler Bedeutung. […]“*

Das IQWiG betrachtet somit zum einen die qualitative Ergebnissicherheit auf Einzelstudien- und Endpunktebene und stuft diese bei randomisierten Studien abhängig vom Verzerrungspotenzial als „hoch“ oder „mäßig“ bzw. bei nicht randomisierten Studien als „gering“ ein. Beim Vorhandensein mehrerer Studien betrachtet das IQWiG bezüglich der darin gezeigten Effekte wiederum verschiedene Aspekte: Homogenität der Ergebnisse, statistische Signifikanz sowie Richtung der Effekte. Sind die Studienergebnisse ausreichend homogen, so wird im Rahmen einer Meta-Analyse ein gemeinsamer Effektschätzer gebildet, der hinsichtlich seiner statistischen Signifikanz überprüft wird. Sind die Studienergebnisse zu heterogen, um sie zu poolen, so wird die Richtung der Effekte betrachtet und in die Kategorien „deutlich

gleichgerichtet“, „mäßig gleichgerichtet“ oder „nicht gleichgerichtet“ eingestuft. Je nach Ergebnis ergibt sich gemäß der Matrix, die das IQWiG in der Tabelle 2 vorgibt, entweder ein „Beleg“, ein „Hinweis“ oder ein „Anhaltspunkt“ für einen Zusatznutzen.

Die Idee, die Aussagesicherheit abhängig von der Anzahl der Studien, der qualitativen Ergebnissicherheit sowie den gezeigten Effekte zu machen, sind ebenso wie die Orientierung an den GRADE-Kriterien bei der Ausarbeitung der Operationalisierung grundsätzlich nachvollziehbar. Der konkreten Ausgestaltung der Operationalisierung, die im Rahmen der frühen Nutzenbewertung angewendet wird, kann von BAYER aber nicht zugestimmt werden. Die bereits hohen Anforderungen an die Beleglage in den Allgemeinen Methoden 4.0 haben sich im Rahmen der Aktualisierung noch einmal stark verschärft und führen nun dazu, dass im Rahmen der frühen Nutzenbewertung von den meisten Arzneimitteln wohl maximal ein „Anhaltspunkt“ für einen Zusatznutzen erreicht werden kann.

Das IQWiG hält auch in den intendierten Aktualisierungen seiner Allgemeinen Methoden an den definierten Anforderungen fest, dass für den Beleg eines Zusatznutzens regelhaft mindestens zwei Studien mit konsistenten Ergebnissen zu einer Fragestellung vorhanden sein müssen. BAYER sieht dies, z. B. in Hinblick auf onkologische Indikationen, als ein ungeeignetes Konzept an. So stellt es in einigen Indikationsgebieten eher eine Ausnahme als die Regel dar, dass zum Zeitpunkt der Zulassung mehr als eine Studie vorliegt. Als Gründe dafür sind erneut die bereits erwähnten ethischen Aspekte zu nennen, wegen derer Studien, die bereits die Überlegenheit eines Wirkstoffs gezeigt haben, nicht ohne weiteres wiederholt werden können. Sowohl aufgrund der niedrigen Prävalenzen und der Schwere mancher Erkrankungen als auch der spezifischen methodischen und ethischen Besonderheiten hinsichtlich der Studienentwicklung für neue z. B. onkologische Therapien, ist die Durchführung mehrerer RCTs häufig weder ethisch vertretbar noch praktisch durchführbar (6). Darüber hinaus gibt es keine regulatorisch formalen Anforderungen oder empirische Evidenz, warum es nötig ist, zwei oder mehr pivotale RCTs durchzuführen, wenn die Wirksamkeit und Überlegenheit eines Wirkstoffes bereits in der ersten Phase-III-Studie statistisch signifikant und klinisch relevant gezeigt wurde.

In dem noch gültigen Methodenpapier des IQWiG (4.0) wird bereits darauf verwiesen, dass ein Nutzenbeleg im Ausnahmefall auch durch eine einzige Studie ableitbar ist, sofern bestimmte Anforderungen erfüllt sind (9). Hierbei wird auf das Dokument „European Medicines Agency. Points to consider on application with: 1. meta-analyses; 2. one pivotal study” (10) verwiesen. Hier wird deutlich ausgeführt, dass bei Berücksichtigung nur einer Studie für den Nachweis eines Nutzenbelegs hohe Anforderungen sowohl an die interne und externe Validität, die Datenqualität als auch an das daraus resultierende Verzerrungspotential gestellt werden. Wann das IQWiG diese Anforderungen als erfüllt betrachtet, ist aber unklar.

Als positiv erachtet BAYER, dass das IQWiG – anders als bei der Ableitung des Ausmaßes des Zusatznutzens aus einer einzelnen statistischen Größe – in seinen Ausführungen angibt, dass

in begründeten Fällen von der regelhaften Operationalisierung abgewichen werden kann. Wenn weitere relevante Faktoren die Einschätzung beeinflussen, ist eine Verringerung oder Erhöhung der Aussagesicherheit möglich. Gleichzeitig wird in diesem Zusammenhang aber befürchtet, dass das IQWiG eher zum Abstufen der Aussagesicherheit neigen wird als zum Aufwerten. So verweist das IQWiG zwar auf die GRADE-Kriterien, die zu einer Abwertung der Ergebnissicherheit führen, vernachlässigt aber jene Kriterien, die konsequenterweise eine Aufwertung nach sich ziehen müssten. Diese selektive Übernahme internationaler Standards ist methodisch nicht nachvollziehbar und wird von BAYER vor allem vor dem Hintergrund, dass die Anforderungen bereits extrem konservativ sind, als problematisch empfunden. Vor allem im Kontext besonders schwerer, wie z. B. onkologischer Indikationen sollte in Abwägung der bereits aufgeführten Punkte und der Spezifika onkologischer Erkrankungen eine differenziertere Ableitung hinsichtlich der Beleglage erfolgen und ein „Beleg“ für einen Zusatznutzens nicht auf Basis des erstellten Anforderungsschemas verwehrt bleiben.

Literatur

1. G-BA. Zusammenfassende Dokumentation über die Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V: Abirateronacetat, vom 29. März 2012. Gemeinsamer Bundesausschuss; 2012; Available from: http://www.g-ba.de/downloads/40-268-1951/2012-03-29_AM-RL-XII_Abirateron_ZD.pdf.
2. IQWiG. Ticagrelor - Nutzenbewertung gemäß § 35a SGB V. IQWiG-Berichte - Jahr 2011 Nr. 96. Auftrag: A11-02. Version: 1.0. Stand: 29.09.2011. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen; 2011; Available from: http://www.g-ba.de/downloads/92-975-5/2011-01-01-D-001_Ticagrelor_IQWiG-Nutzenbewertung.pdf.
3. G-BA. Zusammenfassende Dokumentation über die Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V: Ticagrelor, vom 15. Dezember 2011. Gemeinsamer Bundesausschuss; 2011; Available from: http://www.g-ba.de/downloads/40-268-1826/2011-12-15_AM-RL-XII_Ticagrelor_ZD.pdf.
4. Röhmel J. Gutachten zum Vorschlag des IQWiG zur Bewertung des Ausmaßes des Zusatznutzens im Rahmen der Nutzenbewertung von Arzneimitteln nach § 35a SGB V. Bremen, 21.05.2012. 2012.
5. Djulbegovic B, Kumar A, Soares HP, Hozo I, Bepler G, Clarke M, et al. Treatment Success in Cancer - New Cancer treatment Success identified in Phase 3 Randomized Controlled Trials conducted by the National Cancer Institute-sponsored Cooperative Oncology Groups, 1955 to 2006. Arch Intern Med. 2008;168(6):632-42.
6. Aidelsburger P, Wasem J. Kosten-Nutzen-Bewertungen von onkologischen Therapien. Gutachten für die Deutsche Krebsgesellschaft e.V.; 2008; Available from: http://www.krebsgesellschaft.de/download/gutachten_2.pdf.

7. o.V. Suche nach dem Maß des Gesamtnutzens - Interview mit Prof. Dr. Axel Mühlbacher. Monitor Versorgungsforschung. 2012; 4: 12-16. 2012; Available from: http://www.monitor-versorgungsforschung.de/bilder/kongress-endpunkt-bilder/interview_prof-muehlbacher.

8. G-BA. Zusammenfassende Dokumentation über die Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V: Cabazitaxel, vom 29. März 2012. Gemeinsamer Bundesausschuss; 2012; Available from: http://www.g-ba.de/downloads/40-268-1974/2012-03-29_AM-RL-XII_Cabazitaxel_ZD.pdf.

9. IQWiG. Allgemeine Methoden Version 4.0. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen; 2011; Available from: https://www.iqwig.de/download/IQWiG_Methoden_Version_4_0.pdf.

10. CPMP. Points to consider on application with: 1. meta-analyses; 2. one pivotal study. London: European Medicines Agency (EMA); 2001; Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf.

A.1.4 – Bristol-Myers Squibb GmbH & Co. KGaA

Kommentierung zur
„Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“, Entwurf vom 18.04.2013

Am 19.04.2013 hat das IQWiG seinen Entwurf einer Aktualisierung der Methoden (Version 4.1.) zur Diskussion gestellt. Die Kommentierungsfrist endet am 22.05.2013. Bristol-Myers Squibb möchte die Möglichkeit nutzen, zur Weiterentwicklung der Methoden beizutragen und deshalb einige ausgewählte Punkte kommentieren. Die Stellungnahme gliedert sich in einen ersten Teil, der mehr grundsätzliche bzw. übergeordnete methodische Punkte adressiert (Punkte 1-3) sowie einen zweiten Teil (Punkt 4), in dem einzelne Detailpunkte der Methoden kommentiert werden.

1.) Unterstützung von Entscheidungen des Gemeinsamen Bundesausschusses durch das IQWiG

In der Regel werden Entscheidungen des Gemeinsamen Bundesausschusses (G-BA) durch die IQWiG-Berichte als wesentliche Grundlage unterstützt. Die Berichte des IQWiG sollen den G-BA einerseits zu allen relevanten Bereichen –auch quantitativ – informieren ohne ihm andererseits die Entscheidung vorwegzunehmen. Viele Anmerkungen in dieser Kommentierung basieren auf diesem Verständnis von „Entscheidungsunterstützung“.
Am wichtigsten erscheint uns hier die grundsätzliche Position, dass ein statischer, eindimensionaler Grenzwert für Entscheidungen abzulehnen ist (siehe z.B. Ubel 2003). Ein gutes Beispiel stellen Länder dar (z.B. Großbritannien), in denen vermeintlich ein „Grenzwert“ für die Erstattungsfähigkeit von Gesundheitsleistungen besteht, letztlich aber gilt auch dort eine stochastische Beziehung zur getroffenen Erstattungsentscheidung (Rothgang 2004, Claxton 2013). Über unterschiedliche Systeme hinweg sehen wir, dass die Entscheidungsfindung im Rahmen eines *Health-Technology-Assessments* (HTA) letztlich multidimensional erfolgt, so dass je nach Kriterien des Entscheidungsträgers - selbst im Falle grundsätzlich ähnlicher Bewertungs-Methodik und bei gleicher Datenlage - unterschiedliche Entscheidungen getroffen werden können (z.B. Clement 2009).
Empirisch findet sich europaweit eine Priorität für die Dimensionen Effektivität, Sicherheit und Kosten (Stephens 2013). Die Möglichkeit einer darüber hinaus gehenden, multidimensionalen Berücksichtigung von Kriterien bei der Entscheidungsfindung, z.B. auch von Krankheitslast oder gesellschaftlichen Präferenzen halten wir für wichtig.

Schlussfolgerung: An die Methodik des IQWiG besteht die grundsätzliche Anforderung, einerseits eine nachvollziehbare (und ggf. quantitative) Datenlage für informierte Entscheidungen zu liefern ohne andererseits durch statische, eindimensionale Grenzwerte die multidimensionale Entscheidung des G-BA vorab einzuschränken.

2.) Quantitative Entscheidungskriterien zum Ausmaß eines Effektes

In der Aktualisierung der Methoden des IQWiG findet sich in Tabelle NT1 auf S. 18 (bzw. erläuternd im Anhang S. 26ff) eine Festlegung von **festen Schwellenwerten** für die Feststellung des Ausmaßes eines Effektes. Während eine solche Kategorisierung vor dem Hintergrund der in Punkt 1 dargestellten Entscheidungsunterstützung prinzipiell hilfreich sein kann, so hat der hier vorgeschlagene, spezifische Ansatz einige grundsätzliche Schwächen. Insbesondere beruht der Ansatz ausschließlich auf den **relativen Risiken** der klinischen Effektivität. Hiermit wird ein nur scheinbar neutrales Maß vorgestellt, das jedoch nicht geeignet ist, eine Entscheidungsfindung zu unterstützen. Insbesondere bestehen neben den im Anhang des IQWiG-Entwurfs diskutierten Punkte zusätzliche Abhängigkeiten von:

- der **absoluten Effektgröße** (siehe die bekannte Diskussion um relative und absolute Risikoreduktion).
- Der **Studienanzahl** bzw. **Patientenzahl:** alle Darstellungen und Herleitungen des IQWiG berücksichtigen nicht den Effekt unterschiedlicher Studienanzahl.
- Darüber hinaus sind aus statistischen Gründen für eine wahrscheinliche Erreichung einer Zielgrößenkategorie, die als **Grenze eines Konfidenzintervalls** definiert ist, eine deutliche **Erhöhung der Fallzahl** im Vergleich zu einer üblichen Hypothesentestung notwendig. Die übliche statistische Fallzahlplanung für klinische Studien würde damit regelhaft zu einer zu geringen statistischen Power der Studie für die IQWiG-Kategorien führen. Damit wird die Erreichung der jeweiligen Kategorie zufällig, was nicht im Sinne des Methodenvorschlags sein kann.
- „Anker Mortalität“: nur dieser ist begründet, für alle weiteren Grenzen und die damit **implizierten Werturteile** liegen keine nachvollziehbare Begründung vor. Zusätzlich ist der verwendete „Anker“ nur für die **Onkologie** gültig, da sich ausschließlich auf diesen Bereich die zu Grunde gelegte Publikation bezieht.

Letztgenannter Punkt wiegt u.E. deshalb besonders schwer, da es sich nicht nur um „technische Aspekte“, sondern um prinzipiell die Entscheidungsunterstützung verzerrende Faktoren handelt.

Für besonders kritisch halten wir, dass den Grenzen der anderen Zielgrößenkategorien im Verhältnis zu Mortalität **klare Werturteile** zugrunde liegen. Diese erfolgen aber „statistisch“ implizit und sind quantitativ (und selbst qualitativ) weder transparent noch nachvollziehbar dargestellt oder begründet. Damit besteht insgesamt ein Verzerrungspotential in der Bewertung zwischen verschiedenen Therapiegebieten, das intransparent und nicht quantifizierbar ist.

Dadurch ist die Validität des Konzeptes in allen Therapiebereichen, in denen die Mortalität nicht im Vordergrund steht, fraglich. Dies wird umso deutlicher, wenn man bedenkt, dass die zu Grunde gelegte Untersuchung zur Mortalität in der Onkologie nur durch eine einzige Publikation belegt ist (Djulbegovic 2008). Im Interesse einer möglichst effizienten Verwendung von Mitteln besteht hierdurch ein hohes Verzerrungsrisiko für Entscheidungen; erst kürzlich konnten für quantitative ökonomische Grenzwerte derartige Unterschiede für Therapiegebiete belegt werden (Claxton 2013).

Das quantitative Verhältnis von Mortalität zu anderen Zielgrößenkategorien ist weder aus der bisherigen internationalen Entscheidungspraxis noch aus Präferenzstudien klar bekannt (siehe z.B. Clement 2009, Gyrd-Hansen 2007, Guindo 2012). Bei einem unvalidierten Einsatz der vorgeschlagenen Methodik über alle Therapiegebiete hinweg befürchten wir, dass es zu einer Ungleichbehandlung einzelner Therapiegebiete sowie von Mortalität versus anderen Zielgrößenkategorien kommt. Die vom IQWiG vorgeschlagene Methodik ist neu (Erstveröffentlichung 09/2011) und bislang wissenschaftlich nur unvollständig evaluiert.

Schlussfolgerungen: Die Einflussfaktoren auf die quantitativen Ergebnisse müssen klarer beschrieben werden (z.B. absolute Effektgröße, Studienanzahl), damit nachvollziehbare und keine zufälligen Ergebnisse erhalten werden. Insbesondere fordern wir, dass das quantitative Verhältnis der Zielgrößenkategorien zueinander spezifisch für das jeweilige Therapiegebiet ermittelt oder zumindest validiert wird, um das Verzerrungspotential zu reduzieren. Insgesamt sollte vor einem Einsatz der Methodik eine regelhaft vorgesehene, angemessen wissenschaftliche Begleitung und Evaluation durchgeführt werden, die auch klinisch-medizinische Expertise umfasst.

3.) Besonderheiten der frühen Nutzenbewertung nach §35a SGB V

Zu einem frühen Zeitpunkt der Bewertung von Arzneimitteln, nämlich beim Inverkehrbringen, liegt die Sondersituation vor, dass die Evidenzlage regelhaft auf ≤ 2 randomisierte kontrollierte Studien (RCT) beschränkt ist. In Tabelle 2 auf S. 11 der Methoden 4.1 wird allgemein die Anforderung an die Aussagesicherheit vorgegeben, was entsprechend auch für die frühe Nutzenbewertung nach §35a SGB V von Arzneimitteln gilt.

Hieraus ergibt sich, dass im Rahmen der frühen Nutzenbewertung nach §35a SGB V regelhaft ein „Beleg" nicht erbracht werden kann, da die vorgeschlagenen Anforderungen für ≥ 2 Studien von einer anderen Ausgangssituation ausgehen. Für den „Beleg" seien nicht nur mindestens zwei Studien, sondern zwei „deutlich gleichgerichtete" Studien notwendig. Dies bedeutet, dass beide ein statistisch signifikantes Ergebnis zeigen müssen, was z.B. aufgrund einer Fallzahlplanung für eine gemeinsame Auswertung vorliegen kann, üblicherweise aber nicht vorliegen muss. In diesem Zusammenhang wird insbesondere von der EMA lediglich „*some studies clearly positive*" verlangt (EMA 2001).

In einigen Therapiebereichen wird sogar häufig nur **eine Studie** vorliegen, z.B. aufgrund besonderer Größe oder Population. Insbesondere in der Onkologie sind ethische Vorbehalte gegen die Durchführung einer zweiten Studie abzuwägen. Die entsprechenden Anforderungen werden in den „*Points to Consider*" der EMA (2001) festgelegt und von den europäischen Zulassungsbehörden als Grundlage ihrer Entscheidung zur Anerkennung nur einer pivotalen Studie herangezogen.

Im Rahmen der Nutzenbewertung von Arzneimitteln zum Zeitpunkt des Inverkehrbringens ist deshalb typischer Weise mit der vorgeschlagenen Methodik 4.1 des IQWiG von nicht mehr als einem Hinweis (oder Anhaltspunkt) auszugehen. Während dies klar die bestehende Unsicherheit der Evidenz beschreibt, so kann dies auch als grundsätzliche Schwäche der frühen Nutzenbewertung missverstanden werden. Um dies zu vermeiden und der besonderen Situation einer frühen Bewertung mit eingeschränkter Evidenzlage gerecht zu werden, könnte beispielsweise die Kategorisierung der Aussagesicherheit für die frühe Nutzenbewertung für n=2 und n≥3 gesondert vorgenommen werden. In Tabelle 2 sollte darüber hinaus auf die im Text genannte Möglichkeit verwiesen werden, dass auch mit einer Studie in begründeten Ausnahmefällen eine erhöhte Sicherheit („Beleg") möglich ist.

Schlussfolgerungen: Bei der frühen Nutzenbewertung nach §35a SGB V liegt regelhaft nur eine geringe Studienanzahl (n≤2) vor, so dass sich gemäß vorgeschlagener Methodik 4.1 bestenfalls ein „Hinweis" oder „Anhaltspunkt" ableitet. Eine Anpassung der Aussagesicherheit in der frühen Nutzenbewertung zu einem frühen Zeitpunkt sollte deshalb erfolgen. Dies gilt insbesondere für den Bereich Onkologie, wo häufig nur eine Studie vorliegt.

4.) Kommentierung im Detail:

S. 9, gerichtete Studien: *„Gesamtgewicht dieser Studien ist >80%. Mindestens 2 dieser Studien zeigen statistisch signifikante Effekte*."

Es fehlt eine Begründung für diese Aussagen. Da die Frage des gleichgerichteten Effekts sehr bedeutsam für die Ergebnissicherheit ist, sollte eine Begründung der vorgeschlagenen Operationalisierung erfolgen.

S. 9/10 zu „*mäßig*" bzw. „*deutlich*" gleichgerichteten Effekten:

- Im Methodenpapier 4.1 sind für 2 Studien jeweils für beide Studien signifikante Ergebnisse gefordert, bei mehr Studien für „*mindestens 2*" der Studien (S. 10). Damit werden für wenige Studien besonders hohe Anforderungen an einen „*gleichgerichteten*" Effekt gestellt. Eine Begründung für die unterschiedliche Festlegung je nach unterschiedlicher Studienzahl wird nicht gegeben. Die besonders hohen Anforderungen an Effektgrößen sind insbesondere für die frühe Nutzenbewertung relevant, siehe Kommentierungspunkt 3.
- Auf S. 10 wird das Prädiktionsintervall ab 4 Studien verwendet, um gleichgerichtete Effekte zu operationalisieren. Das Prädiktionsintervall ist eine derzeit noch wenig angewandte Methode, für deren Einsatz zur Operationalisierung von gleichgerichteten Effekten nur sehr wenig Erfahrung vorliegt. Eine angemessene wissenschaftliche Fundierung sollte deshalb erfolgen.

Werturteile S. 10/11:

- Auf S.10 *„welche Abweichungen im Design zwischen Studien noch akzeptabel sind, hängt von der Fragestellung ab*". Mit dieser Aussage erfolgt ein Werturteil, das klar als solches kenntlich zu machen ist und begründet werden muss. Eine systematische transparente Evaluation unter Einbeziehung externer Experten (s. Kommentierungspunkt 2) wäre auch hier wünschenswert.
- Auch auf S. 10/11 *„begründete Zweifel an der Übertragbarkeit auf die Behandlungssituation in Deutschland können z.B. zu einer Verringerung…*" erfolgt ein Werturteil, das transparent zu machen und zu begründen ist. Da hier Evidenz in ihrer Aussage herabgesetzt oder sogar ausgeschlossen wird, sind besondere Anforderungen insbesondere an den Nachweis der Nicht-Übertragbarkeit für Deutschland zu stellen.
  Eine gleichzeitige hohe externe Validität ist im Falle von RCTs mit hoher interner Validität nur schwer umzusetzen, vor allem wenn hohe Fallzahlen notwendig sind. Im Fall der Nutzenbewertung nach §35a SGB V liegt der Fokus bisher auf der internen Validität – selbstverständlich bei akzeptabler externer Validität für Europa, aber auch unter Berücksichtigung der Praktikabilität. In Zeiten einer internationalisierten Medizin und regelhaft internationalen Studiendurchführungen darf damit generierte, weltweit gültige Evidenz nicht „willkürlich" in ihrer Aussagesicherheit für Deutschland abweichend bewertet werden. Es sei darauf hingewiesen, dass sich auch innerhalb Deutschlands in vielen Bereichen der Medizin regional sehr unterschiedliche Versorgungssituationen finden. Es besteht deshalb die Gefahr, dass das Argument „externe Validität" als Argument für eine Nichtberücksichtigung von Evidenz „pauschal" zur Anwendung kommen könnte.
- Wir schlagen deshalb vor, den Abschnitt hinsichtlich der Begründung eines Ausschlusses im Sinne des beabsichtigten Vorgehens zu präzisieren, beispielsweise zu ergänzen: „Eine Verringerung oder Erhöhung der Aussagesicherheit erfolgt nur im Ausnahmefall und mit ausführlicher Begründung."

S. 11, Satz oben *„Auf der anderen Seite können z.B. große Effekte…*" – Dieser Satz ist unverständlich, ggf. fehlerhaft formuliert: wieso erhöht eine „eindeutige Richtung eines vorhandenen Verzerrungspotentials" die Sicherheit der Aussage?

S. 12, erster Punkt zum Gewicht der Studie: „*zwischen 25 und 75%*“. Es fehlt eine Quellenangabe bzw. Begründung für das Intervall. Wieso nicht z.B. zwischen 20% und 80%? Zudem sollte deutlich gemacht werden, ob „zwischen“ die Grenzen inkludiert oder nicht.

S: 13 unten „*alternative Verfahren der multikriteriellen Entscheidungsfindung oder der Präferenzerhebung*“:
Die entsprechende Abwägung und Methodendiskussion muss im jeweiligen Einzelfall getroffen werden. Eine generelle Bevorzugung einer Methode bzw. grundsätzlicher Ausschluss des QALY Konzeptes („*Probleme gerade der häufig verwendeten QALYs […] sollten alternative Verfahren*“) widerspricht der neutralen Wahl des jeweils am besten passenden Konzeptes.
Hochqualitative multikriterielle Entscheidungsfindungen sind methodisch sehr komplex (Bridges 2011, Johnson 2013) und entsprechend mit hohem Zeit- und Kostenaufwand verbunden. Es bestehen insbesondere auch hier viele Fehlermöglichkeiten, die zu verzerrten Aussagen führen können. Eine grundsätzliche Präferenz für derartige Methoden kann deshalb zwar methodisch gerechtfertigt sein, muss aber gegenüber den bestehenden Umfang an praktischer Erfahrung (z.B. mit dem QALY Konzept) individuell abgewogen werden.
Es wird deshalb empfohlen, die Formulierung abzuschwächen, z.B. „… sollten *sofern verfügbar* alternative Verfahren der multikriteriellen Entscheidungsfindung….“.

S. 15 unten: „*die Bewertung der Kosten auf der Basis der Standards der Gesundheitsökonomie*“. Auch um weiterer Methodenentwicklung in Zukunft Raum zu geben, sollte nicht der Eindruck erweckt werden, die Gesundheitsökonomie beschränke sich auf die derzeitig im Verfahren nach §35a SGB V angewandte, d.h. auf kurzfristige direkte Kosten ausgerichtete Bewertungen. Beispielsweise findet sich im sog. Hannoveraner Konsens eine Darstellung der Breite der gesundheitsökonomischen Möglichkeiten (von der Schulenburg 2007). Formulierungsvorschlag für die derzeitige Bewertung als eine „Bewertung der Kosten auf Basis der Fachinformation“.

S. 17 „relatives Risiko“
Bisher sieht die Formatvorlage des G-BA in Modul 4 die „*Odds Ratio*“ anstelle des relativen Risikos vor. Zu den Vor- und Nachteilen von „Relativem Risiko“ und „Odds Ratio“ besteht bekanntlich eine umfangreiche Literatur. Die Cochrane Collaboration wie auch internationale Literatur – so wie bisher auch vom IQWiG gehandhabt (siehe bisherige Produkte des IQWiG) – sehen keinen klaren Vorteil für das eine oder andere Verfahren (z.B. Deeks 2002).
Wird nun ausschließlich auf das relative Risiko fokussiert, so erhöht dies insbesondere den Analyseaufwand für international ausgewertete Studien, da das relative Risiko obligat „*ggf. selbst berechnet*“ (S. 17) werden muss. Diesem erhöhten Aufwand in der Analyse und Darstellung im Dossier muss ein klarer Vorteil in der Aussagekraft gegenüberstehen. Eine entsprechende Begründung hierfür fehlt.

S . 18, Tabelle NT1
Während eine Begründung für die Zielgrößenkategorie „*Gesamtmortalität*“ bei „*erheblichem Ausmaß*“ vorliegt (zitierte Analyse für die Onkologie), fehlt diese für die abgeleiteten und anderen Kategorien. Hiermit wird eine quantitative Scheingenauigkeit in der Aussage erzeugt, die so nicht vorhanden ist. Eine detaillierte Behandlung dieses sehr bedeutsamen, mehr grundsätzlichen methodischen Punktes erfolgt unter Punkt 2 dieser Kommentierung.

S 19, letzter Absatz: „*Gesamtaussage zum Ausmaß des Zusatznutzens bei gemeinsamer Betrachtung aller Endpunkte, ist eine strenge Formalisierung nicht möglich*“. Dies ist eine wichtige Aussagen, der wir uneingeschränkt zustimmen. Ziel der Bewertung ist vielmehr eine Unterstützung der Entscheidungsfindung (im Sinne „begründeter Vorschlag für eine Gesamtaussage“), siehe

allgemeiner Punkt 1 dieser Kommentierung. Entsprechend dieser Zielsetzung und der bestehenden methodischen Limitationen bezweifeln wir, dass eine strenge und therapiegebietsübergreifende Anwendung der quantitativen Methodik - so wie derzeit vorgeschlagen - sinnvoll möglich ist (siehe Punkt 2 dieser Kommentierung).

S. 21 oben: „*zufällige Effekte*“. Die „Peto-OR“ sollte als „begründeter Ausnahmefall“ erwähnt werden.

S. 21 Einstufung der Heterogenität gemäß I^2
Die Aussage „*in der Regel wird von einer statistischen Zusammenfassung abgesehen, falls der Heterogenitätstest einen p-Wert unter 0,2 liefert*“ steht zumindest teilweise im Widerspruch zu einer früher von Mitarbeitern des Instituts vorgeschlagenen, differenzierteren Methodik (Skipka und Bender, 2010).
Wie korrekt festgestellt wird, ist die Einstufung einer „zu großen“ Heterogenität anhand von I^2 kontextabhängig. Die Vorstellung der Kategorien von I^2 als „grobe Einschätzung“ entspricht dem Verständnis in der Publikation von Higgins und Thompson (2002). Entsprechend sollten diese Kategorien auch umgesetzt werden.

S. 22 „*ökologischer Studien der Epidemiologie [206]*“. Das Zitat ist für die getroffene Aussage zu Meta-Regressionen für uns nicht nachvollziehbar. Wir bitten um eine Klarstellung.

S. 22 Prädiktionsintervall
Die Anwendung von Prädiktionsintervallen findet noch keinen breiten Einsatz, kann aber bei der Interpretation von Meta-Analysen hilfreich sein (Riley 2011, Kriston 2013). Relevant hierbei ist die Aussage des Prädiktionsintervalls zu berücksichtigen, das den Effekt einer einzelnen (neuen) Studie vorhersagt. Dementsprechend dient das Prädiktionsintervall „*nicht zur Beurteilung der statistischen Signifikanz eines Effekts*“ (IQWIG Methoden 4.1. S. 22). Um genau eine solche, unangemessene Interpretation des Prädiktionsintervalls zu vermeiden, sollte genauer beschrieben werden, wofür das Prädiktionsintervall genutzt wird.

Zitat Guddat [N5]. An der zitierten Arbeit waren Mitarbeiter des Instituts beteiligt. Ein „Eigenzitat“ sollte nur eingeschränkt zur Begründung herangezogen werden.

Auf die derzeit noch beschränkte Umsetzung des Prädiktionsintervalls in meta-analytischer Software würden wir gerne hinweisen.

S. 23, Subgruppenanalysen
Es sollte klarer auf die bekannten Limitationen des Interaktionstests hingewiesen werden (Testgüte). Diese hängt u.a. von der Anzahl der Studien ab (Higgins 2002). Entsprechend kann z.B. bei falsch positivem Interaktionstest eine Darstellung eines gemeinsamen Effektschätzers angemessen sein.

S. 26 ff Anhang „*Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens*“
Bis Tabelle NT4 ist das Vorgehen für uns prinzipiell nachvollziehbar, da es nicht-quantitativ bleibt. Wie auf S. 31 festgestellt wird, folgen dann „*notwendigerweise zu treffende*“ Werturteile. Auf S. 32 wird die als Anker dienende Publikation von Djulbegoic [N3] zitiert. Diese Publikation bezieht sich auf die Onkologie und die Zielgröße Gesamtmortalität. Bereits diese Festlegung auf der Basis einer einzelnen Publikation erscheint uns nur begrenzt begründet. Die weiteren, abgeleiteten Kategorien sind jedoch überhaupt nicht begründet. Im allgemeinen Kommentierungspunkt 2 wird auf diese und weitere Probleme des vom Institut gewählten methodischen Ansatzes eingegangen.

Wir schlagen vor, dass zumindest eine therapiegebietsspezifische Anpassung bzw. Überprüfung der Grenzwerte erfolgen sollte. Weiterhin wollen wir darauf hinweisen, dass für die Darstellung auf S. 36 weitere Einflussfaktoren untersucht werden müssten, z.B. Normierung vs. Basisrisiko, Anzahl der Studien (vs. Patientenzahl). Nur so können die getroffenen Werturteile im Sinne einer Entscheidungsunterstützung transparent dargestellt werden. Eine detailliertere, unabhängige wissenschaftliche Evaluation vor einem Einsatz der Methodik zur Entscheidungsunterstützung erscheint uns ratsam.

Literatur

Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, Johnson FR, Mauskopf J. Conjoint analysis applications in health--a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value Health. 2011 Jun;14(4):403-13.

Claxton K, Martin S, Soares M, Rice N, Spackman E, Hinde S, Devlin N, Smith PC, Sculpher M. Methods for the Estimation of the NICE Cost Effectiveness Threshold. 2013. Online verfügbar: http://www.york.ac.uk/media/che/documents/papers/researchpapers/CHERP81_Methods_estimation_NICE_costeffectiveness_threshold.pdf [Link verifiziert 10.05.2013]

Clement FM, Harris A, Li JJ, Yong K, Lee KM, Manns BJ. Using effectiveness and cost-effectiveness to make drug coverage decisions: a comparison of Britain, Australia, and Canada. JAMA. 2009 Oct 7;302(13):1437-43.

Deeks JJ, Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. STATISTICS IN MEDICINE Statist. Med. 2002; 21:1575–1600

Djulbegovic B, Kumar A, Soares HP, Hozo I, Bepler G, Clarke M et al. Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006. Arch Intern Med 2008; 168(6): 632-642.

European Medicines Agency. Points to consider on application with: 1. meta-analyses; 2. one pivotal study. 31.05.2001. Online verfügbar: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf [Link verifiziert 10.05.2013]

Guindo LA, Wagner M, Baltussen R, Rindress D, van Til J, Kind P, Goetghebeur MM. From efficacy to equity: Literature review of decision criteria for resource allocation and healthcare decisionmaking. Cost Eff Resour Alloc. 2012 Jul 18;10(1):9

Gyrd-Hansen D, Kristiansen IS. Preferences for 'life-saving' programmes: small for all or gambling for the prize? Health Econ. 2008 Jun;17(6):709-20.

Higgins JPT und Thompson SG. Quantifying heterogeneity in a meta-analysis. Statist. Med. 2002; 21:1539–1558

Johnson RF, Lancsar E, Marshall D, Kilambi V, Mühlbacher A, Regier DA, Bresnahan BW, Kanninen B, Bridges JF. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value Health. 2013 Jan-Feb;16(1):3-13.

Kriston L. Dealing with clinical heterogeneity in meta-analysis. Assumptions, methods, interpretation. Int. J. Methods Psychiatr. Res. 22(1): 1–15 (2013)

Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011, 342:d549.

Rothgang H, Niebuhr D, Wasem J, Gress S. Das National Institute for Clinical Excellence (NICE) - Staatsmedizinisches Rationierungsinstrument oder Vorbild für die evidenzbasierte Bewertung medizinischer Leistungen? Gesundheitswesen. 2004 May;66(5):303-10.

Skipka G und Bender R. Intervention effects in the case of heterogeneity between three subgroups. Assessment within the framework of systematic reviews. Methods Inf Med. 2010;49(6):613-7.

Stephens J, Handke B, Doshi J, on behalf of the ISPOR HTA SIG Research Methods & Principles Working Group. A SURVEY OF HTA RESEARCH METHODS AND TRENDS IN EUROPE. Poster ISPOR 2013 in Prag. Online verfügbar: http://www.ispor.org/sigs/hta_ebr/hta_grp.asp [Link verifiziert 10.05.2013]

Ubel PA, Hirth RA, Chernew ME, Fendrick AM. What is the price of life and why doesn't it increase at the rate of inflation? Arch Intern Med. 2003 Jul 28;163(14):1637-41. Review.

Von der Schulenburg JM, Greiner W, Jost F, Klusen N, Kubin M, Leidl R, Mittendorf T, et al. Deutsche Empfehlungen zur gesundheitsökonomischen Evaluation – dritte und aktualisierte Fassung des Hannoveraner Konsens. Gesundh ökon Qual manag 2007; 12: 285-290.

A.1.5 – Bundesarbeitsgemeinschaft Selbsthilfe von Menschen mit Behinderung und chronischer Erkrankung und ihren Angehörigen e. V. – BAG Selbsthilfe

Von: GF Geschaeftsfuehrer [redacted]
Gesendet: Dienstag, 21. Mai 2013 10:11
[redacted]

[redacted]

unter Bezugnahme auf Ihre Email vom 22.04. möchte ich Ihnen zum Entwurf des Methodenpapiers 4.0 die Rückmeldung geben, dass auf Seite 14 die Beschreibung des Zusammenhangs bzw. die Abgrenzung im Rahmen von Patientenpräferenzen zur frühen Nutzenbewertung sinnvoll wäre.

Mit freundlichen Grüßen

Dr. Martin Danner
Bundesgeschäftsführer

BAG SELBSTHILFE
Bundesarbeitsgemeinschaft Selbsthilfe von Menschen mit Behinderung
und chronischer Erkrankung und ihren Angehörigen e.V.
Kirchfeldstr. 149, 40215 Düsseldorf
[redacted]

A.1.6 – Bundesverband der Arzneimittel-Hersteller e. V. (BAH)

Bundesverband der Arzneimittel-Hersteller e.V. B.A.H

Stellungnahme

des Bundesverbandes der Arzneimittel-Hersteller e.V. (BAH) zur „Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neuer Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1 (Entwurf 18.4.2013)“ des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)

Der Bundesverband der Arzneimittel-Hersteller e.V. (BAH) vertritt die Interessen seiner 467 Mitglieder, darunter 323 Arzneimittel-Hersteller, gegenüber der Bundesregierung, dem Bundestag, dem Bundesrat, sowie gegenüber Akteuren und Institutionen des Gesundheitswesens. Der BAH ist der mitgliederstärkste Verband im Arzneimittelbereich. Die politische Interessenvertretung und die Betreuung der Mitglieder erstreckt sich zum einen auf den Bereich der Selbstmedikation, zum anderen auf das Gebiet der rezeptpflichtigen Arzneimittel mit Ausnahme der patentgeschützten Präparate.

Der BAH bedankt sich für die Möglichkeit, zu dem benannten Dokument des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) Stellung zu nehmen und möchte durch die Stellungnahme die qualitative Arbeitsweise des IQWiG, die nach § 139a Abs. 4 SGB V in transparenter Weise durchzuführen ist, unterstützen.

Der BAH begrüßt ausdrücklich die Überarbeitung hin zu einer detaillierten Darstellung der Arbeitsweise des IQWiG.

Der BAH nimmt wie folgt Stellung:

ad 2.1.1 Bericht B) Allgemeine Anmerkungen zum Stellungnahmeverfahren (Anhörung)

Ein Hinweis zu der Satzung der Stiftung für Qualität und Wirtschaftlichkeit bei der Nennung des Kuratoriums (Verweis auf § 8 der Satzung) und ein Hinweis zur transparenten Auflistung der Mitglieder des Kuratoriums auf der Website des IQWiG wären wünschenswert.

ad 2.2.3 Review der Produkte des Instituts

Die Beschreibung der Qualitätssicherung wurde dahingehend verändert, dass das externe Qualitätssicherungsverfahren, im Speziellen das externe Reviewverfahren für die Erstellung von Berichten und z.T. auch von Gesundheitsinformationen als externes Qualitätssicherungsverfahren (S.7, zweiter Absatz), nur noch einer „kann“-Bestimmung unterliegt. Um einen hohen Qualitätsstandard des IQWiG zu gewährleisten, ist eine externe

Begutachtung der Arbeitsschritte stets notwendig, so dass der zweite Absatz der Seite 7 in seine Ursprungsform zurückgesetzt werden sollte.

Nicht verständlich ist, warum eine Veröffentlichung der externen Reviews nicht erfolgt (S. 7, 4. Abschnitt). Dies widerspricht dem Anspruch des IQWiG auf eine transparente Arbeitsweise (nach § 139a Abs. 4 SGB V). Eine Offenlegung der sachbezogenen Inhalte ist notwendig.

Weitergehend sollte der Name von externen Sachverständigen nicht „*i.d.R.*“ (Seite 7, 4. Abschnitt) veröffentlicht werden, sondern „*stets“*. Dies ist notwendig, um die Transparenz (nach § 139a Abs. 4 SGB V) zu gewährleisten und eine reliable Qualität der Arbeitsweise des externen Sachverständigen und damit auch des IQWiG darstellen zu können. Die ohne Ausnahme praktizierte Namensnennung der externen Sachverständigen würde eine öffentliche Kontrolle von Interessenskonflikten herbeiführen und damit dem § 139b Abs.3 S. 2 SGB V Rechnung tragen. Mögliche Interessenskonflikte können sicherer erkannt und die fachliche Unabhängigkeit der Sachverständigen durch die Öffentlichkeit kontrolliert werden.

ad 3.1.4 Endpunktbezogene Bewertung

Wie auch in weiteren Kapiteln (z.B. 3.1.5 Zusammenfassende Bewertung) wird die Terminologie von *„(Zusatz-)Nutzen und Schaden“* verwendet.
Vorteilhaft wäre, sich dem international fachwissenschaftlich üblichen Sprachgebrauch von Nutzen und Risiko anzupassen. Das Ersetzen des Wortes *„Schaden“* durch das Wort *„Risiko“* wäre wünschenswert.
Im allgemeinen Sprachgebrauch suggeriert das Wort *„Schaden“* einen möglichen juristischen Haftungsfall und ist im „Sinne einer unfreiwilligen Einbuße“[1] zu deuten. Dies entspricht z.B. nicht den unerwünschten Arzneimittelwirkungen, die in der endpunktbezogenen Bewertung eingebunden werden. Auch ist in der medizinischen Fachsprache in Deutschland, exemplarisch belegt durch die (Muster-) Berufsordnung (Stand 2011) der Bundesärztekammer in § 6 „Mitteilung von unerwünschten Arzneimittelwirkungen“[2], nicht die Terminologie von Schaden in Bezug auf Arzneimittel vorrangig in Verwendung.

In diesem Kapitel sowie in weiteren Kapiteln (z.B. 3.1.5 Zusammenfassende Bewertung) wird neben der Terminologie *„Aussagesicherheit“* z.B. der Begriff *„Beleglage“* (z.B. Seite 9 *„Bei der Ableitung der Beleglage für einen Endpunkt sind die Anzahl der vorhandenen Studien….“*) verwendet. Um die Verständlichkeit zu verbessern, wird empfohlen, durchgängig von „Aussagesicherheit“ zu sprechen, zumal es bei der Verwendung von *„Beleglage“* zu Irritationen mit der Aussagesicherheits-Kategorie „Beleg“ geben kann.

Im ersten Satz auf Seite 9 *„Das Institut verwendet die … Ausmaßes der qualitativen Ergebnissicherheit auf Einzelstudien- und Endpunktebene.“* erscheint der Begriff „Endpunktebene“ nicht passend, da die nachfolgende Aufzählung sich lediglich auf Einzelstudienebene bezieht.

In dem Satz *„Bei der Ableitung der Beleglage für einen Endpunkt sind die Anzahl der vorhandenen Studien, deren qualitative Ergebnissicherheit sowie die in den Studien gefundenen Effekte von zentraler Bedeutung.“* (Seite 9, 2. Abschnitt) könnte das Maß der

[1] Bundeszentrale für politische Bildung zitiert nach http://www.bpb.de/nachschlagen/lexika/recht-a-z/22822/schadensersatz am 14.5.2013
[2] Bundesärztekammer zitiert nach http://www.bundesaerztekammer.de/page.asp?his=1.100.1143 am 14.5.2013

internen und externen Variabilitäten integriert werden. Dieser Abschnitt sollte die Überschrift „qualitative Ergebnissicherheit auf Endpunkteebene“ erhalten.

Um dem Transparenzgebot des IQWiG Rechnung zu tragen, sollten Literaturstellen, die als Quellen herangezogen werden, eingearbeitet bzw. zitiert werden. Der Leser bleibt z.B. in Ungewissheit, ob der zweite Aufzählungspart auf Seite 9 wissenschaftlich anerkannte Inhalte wiedergibt.

Es fällt auf, dass das neu aufgenommene *„Prädiktionsintervall“* starke Betonung und Einbindung erhält und andere Methoden, wie z.B. I^2 in die Gefahr der Bedeutungslosigkeit gelangen.

Die Aufzählungspunkte auf Seite 10 erscheinen nicht kongruent mit der Seite 9: Der Aufzählungspunkt *„2 Studien: Gleichgerichtete Effekte sind immer deutlich gleichgerichtet“* sollte entsprechend dem Punkt auf Seite 9 *„mindestens 2 dieser Studien zeigen statistisch signifikante Ergebnisse“* um den Zusatz *„Beide Studien weisen statistisch signifikante Ergebnisse auf“* ergänzt werden.
Das Fehlen der Literaturquellen lässt den Leser in Unsicherheit, ob die Darstellung eines Prädiktionsintervalls, falls mindestens 4 Studien vorliegen, korrekt ist. Es mag notwendig sein, dass die Beurteilung der Einstufung in mäßig gleichgerichtete und deutlich gleichgerichtete Effekte auch bei Vorliegen von 4 Studien ohne Prädiktionsintervall vorzunehmen ist. Eine Ergänzung der Beschreibung der Arbeitsweise wäre wünschenswert.

Um die Entscheidungsfindung des IQWiG transparent und reproduzierbar zu machen, ist die Formulierung *„In der Regel…“* (Seite 10, letzter Abschnitt) zu erweitern mit der Definition der Ausnahmen.

Mittels des Satzes *„Welche Abweichungen im Design zwischen Studien noch akzeptabel sind, hängt von der Fragestellung ab.“* hat der Leser keine adäquate Beschreibung erhalten und eine Beliebigkeit könnte unterstellt werden. Aufgrund der bereits vielfältigen Erfahrungen des IQWiG mögen Konkretisierungen, Auflistungen von Beispielen für diesen Fall möglich sein. Dies ist auch im Hinblick auf die Formulierung *„In begründeten Fällen beeinflussen weitere Faktoren diese Einschätzung“* bereits angemessen vorgenommen worden.

Die Beschriftung der Tabelle 2 hin zu „Anforderungen der Aussagesicherheiten beim Vorliegen von Studien dergleichen qualitativen Ergebnissicherheit“ könnte zur Verbesserung der Verständlichkeit dienlich sein.
Der Hinweis a der Tabelle 2 *„zur Erläuterung des Begriffs: siehe Text“* ist für den Leser keine Hilfe, eine Konkretisierung u.a. mit Nennung der Seitenzahl wird empfohlen.
Einen Beleg für das Vorgehen des zweiten Aufzählungspunktes *„…wobei die Aussagesicherheit durch die minimale qualitative Ergebnissicherheit aller einbezogenen Studien bestimmt wird“ ist aufzuführen. Unter dem Begriff „höherwertige Studien“* kann der Leser keine eindeutige Zuordnung der qualitativen Ergebnissicherheit treffen. Die konsistente Verwendung der vom IQWiG aufgestellten/definierten Begriffe ist notwendig, um den inhaltlichen Blick ausrichten zu können. Nicht nachvollziehbar ist die Vorgabe *„…, wobei die Aussagesicherheit durch die minimale qualitative Ergebnissicherheit aller einbezogenen Studien bestimmt wird.“.* Vorab wurde beschrieben, dass die Aussagesicherheit von den Kriterien *„Anzahl der Studien“, „qualitative Ergebnissicherheit“* und *„Effekte“* bestimmt wird, eine Herleitung, warum nur die minimalste qualitative Ergebnissicherheit für die Aussagesicherheit Verwendung findet, fehlt und erscheint im Sinne der Evidenzbasierten Medizin nicht korrekt.

ad 3.1.5 Zusammenfassende Bewertung

Erneut (siehe auch ad 3.1.4 Endpunktbezogene Bewertung) wird auf die Vorteile der Verwendung des Begriffs „Risiko“ anstelle des Begriffs „Schaden“ hingewiesen: Zum einen ist die internationale Fachsprache entsprechend ausgerichtet und zum anderen ebenso –wie exemplarisch gezeigt- die deutsche Sprache. Ein Beitrag zur besseren Verständlichkeit wäre somit erzielbar.

Der BAH begrüßt die aufgeführte gleichzeitige Würdigung von Nutzen und „Schaden“ mit dem dargestellten Verfahren der Analytic Hierarchy Process (AHP) und Conjoint-Analyse (CA). Der BAH betont die vorrangige Positionierung der Patientinnen und Patienten sowie der behandelnden Ärztinnen und Ärzte in diesen Verfahren. Es bleibt unklar, wie die praktische Umsetzung unter der Einbeziehung der Stärken und Schwächen beider Verfahren stattfinden kann. Eine umfassende und transparente Erörterung, besonders mit dem Schwerpunkt der Patientschaft und der klinisch-ethischen Fachexpertisen ist notwendig. Der BAH empfiehlt daher nachdrücklich eine Präzisierung der Arbeitsweise zur zusammenfassenden Bewertung.

Da das IQWiG bereits auf zahlreiche Verfahren zurückgreifen kann, wäre eine Konkretisierung der qualitativen oder semiquantitativen Abwägungen vorteilhaft.

ad 3.3.3 Nutzenbewertung von Arzneimitteln gemäß §35a SGB V

Der Leser wird durch inkonsistente oder fehlerhafte Verwendungen von Begriffen und strukturdefizitäre Inhalte irritiert.
Es fällt auf, dass die Terminologie z.B. des Kapitels „3.1.4 Endpunktbezogene Bewertung“ nicht konsistent im hiesigen Kapitel wiederzufinden ist. Der BAH regt zur besseren Verständlichkeit an, die klare Verwendung des Kapitels 3.1.4 mit z.B. dem Begriff *„Aussagesicherheit“* in das hiesige Kapitel einzupflegen. Beispielhaft werden die Worte *„Aussagekraft“* (S. 15, 2. Abschnitt), *„Ableitung zur Aussage von Beleglage“* (S.16, 1. Aufzählung) oder *„Wahrscheinlichkeit“* (S.16, 1. Aufzählung) anstelle des Begriffs *„Aussagesicherheit“* verwendet.

Im 2. Schritt (S.16) wird das *„Ausmaß der Effektstärke“* mit *„quantitativer Aussage“* gleichgesetzt. Dies erscheint problematisch, da es sich um ordinale Aussagekategorien handelt, die im 2. Schritt eingeführt werden. Der BAH schlägt zur besseren Verständlichkeit vor, anstelle *„quantitative Aussagen“* den Begriff *„Ausmaßkategorien“* bzw. *„Ausmaßkategorien des Effektes“* stets zu verwenden. In der Tabelle NT1 wird z.B. der erste vorgeschlagene Begriff in der ersten Spalte bereits verwendet.

Mit weiterem Blick auf die Tabelle NT1 (S. 18) wird leider die unzureichende Ausarbeitung des Kapitels erneut deutlich: Die Überschrift der Tabelle NT1 „Schwellenwert zur Feststellung des Ausmaßes eines Effektes“ ist fehlerhaft, da nicht durch die Schwellenwerte das Ausmaß eines Effektes festgestellt wird, sondern die aufgeführten Schwellenwerte zur Definition der Ausmaßkategorie des Effektes herangezogen werden.
Auch sollte die Überschrift der Tabelle NT1 identisch mit der Überschrift der Tabelle NT5 sein, da diese Tabellen inhaltlich gleich sind.

Die aufgeführten Schwellenwerte zeigen eine Hierarchisierung der Endpunkte auf, wobei *„die Schwellenwerte sind …umso größer (im Sinne näher an 1), je schwerwiegender das Ereignis ist.“* (S. 17, letzter Abschnitt). Diese Schwellenwerte beziehen sich einzig auf Studienergebnisse. Die klinische Erfahrung und Expertise der behandelnden Ärztinnen

und Ärzte sowie die Erfahrungen und Präferenzen der betroffenen Patientinnen und Patienten sind unabdingbar in die Erstellung der Ausmaßkategorie bei einer Arbeitsweise der Evidenz-basierten Medizin einzubeziehen. Es erscheint ethisch und auch im Sinne der Evidenz-basierten Medizin nicht vertretbar, solche Schwellenwerte festzulegen und einzig zur Kategorisierung des Effektausmaßes in Abhängigkeit der Zielgrößenkategorien heranzuziehen. Eine transparente Erarbeitung der Rahmenbedingungen mit vom IQWiG unabhängigen klinischen, ethischen, biometrischen Fachexpertisen sowie Patientenvertretung zur Operationalisierung der Kategorisierung von Effektausmaßen ist erforderlich.

Das Wort „Rationale“ (Seite 16) wird vom aktuellen Duden nicht in diesem Kontext geführt. Um die Verständlichkeit zu optimieren, wäre „Begründung“ eher passend.

„Das grundsätzliche Konzept sieht vor, für relative Effektmaße Schwellenwerte für Konfidenzintervalle in Abhängigkeit von anzustrebenden Effekten abzuleiten, die wiederum von Qualität der Zielgröße und den Ausmaßkategorien abhängen.“
Durch den Versuch der Komprimierung eines komplexen Sachverhaltes, der im Anhang verständlich dargestellt ist, sind Inhalte verloren gegangen. Besonders wichtig erscheint die Erläuterung der *„Abhängigkeit von anzustrebenden Effekten“*. Wie werden die anzustrebenden Effekte festgelegt? Dies sollte nicht in einem Anhang aufgeführt werden, sondern in das hiesige Kapitel integriert werden. Ebenso wichtig erscheint die Korrektur der Ausdrucksform *„Qualität der Zielgröße“* (s. auch Seite 17f). Diese kann als eine deformierende Formulierung -insbesondere aus Patientensicht- empfunden werden und ist auch nicht inhaltlich notwendig bzw. sogar missverständlich. Der BAH schlägt vor, statt *„Qualität der Zielgröße“* stets den Begriff *„Zielgrößenkategorie“* zu verwenden.

„Vom Effektmaß relatives Risiko ausgehend werden Zähler und Nenner immer so gewählt, dass sich der Effekt (sofern vorhanden) als Wert <1 realisiert. D.h. ein Effekt ist umso stärker, je niedriger der Wert ist ….Daher muss zur Festlegung des Ausmaßes des Effektes für jede binäre Zielgröße anhand inhaltlicher Kriterien unter Berücksichtigung der Art des Endpunktes und der zugrunde liegenden Erkrankung entschieden werden, welches Risiko betrachtet wird - das für das Ereignis oder das für das Gegenereignis.“ (S.17)
Es ist begrüßenswert, dass man positive und negative Effekte eines Arzneimittels in der Kategorisierung gleich methodisch behandeln möchte, jedoch muss der Aufwand für die Praxis bei der Ergebnisdarstellung von Effekten nach der derzeit dargestellten Methodik hinterfragt werden. Dadurch würde deutlich mehr Analyseaufwand/Umrechnungsaufwand (Kehrwertbildung) anfallen und nationale sowie internationale Publikationen in wissenschaftlichen Zeitschriften könnten ggf. nicht mehr direkt Verwendung finden. Weitergehend ist –wie oben bereits erläutert- zu betonen, dass die klinischen Erfahrungen und Expertisen der behandelnden Ärztinnen und Ärzte sowie die Erfahrungen und Präferenzen der betroffenen Patientinnen und Patienten unabdingbar in die Bewertung/Kategorisierung einzubinden sind. Auch dürfen die biometrisch ermittelten Maßzahlen keine Überbewertung im Vorgehen der Evidenz-basierten Medizin erhalten, zumal z.B. große Endpunktstudien in der Regel aus ethischen Erwägungen mit Iterimsauswertungen durchgeführt und ggf. frühzeitig abgebrochen werden.

„Je nach Qualität der Zielgröße muss das Konfidenzintervall vollständig unterhalb eines bestimmten Schwellenwertens liegen, um das Ausmaß als gering, beträchtlich oder erheblich anzusehen. “
Missverständlich könnte gedeutet werden, dass die Ausmaßkategorie definiert wird je nachdem, ob das Konfidenzintervall des Effektes bezogen auf die Zielgröße vollständig unter dem Schwellenwert liegt.
Der BAH schlägt vor: „Für die drei Zielgrößenkategorien werden unterschiedliche Schwellenwerte der jeweiligen Ausmaßkategorie festgelegt (s. Tabelle NT1). Zur Zuordnung eines Effektes in eine Ausmaßkategorie muss stets die obere Grenze des 95%-Konfidenzintervalles kleiner als der definierte Schwellenwert sein.“

Die Struktur z.B. des Abschnittes „A) Binäre Zielgrößen“ wäre logisch mit der Beschreibung der drei Zielgrößenkategorien zu beginnen, dann wäre auf die Berechnung der Effekte und schließlich auf die Schwellenwerte und Zuordnung einzugehen.

Bisher nicht in diesem Abschnitt erläutert wird z.B. der Umgang mit Odds Ratios. (Der Anhang, in dem dies erklärt wird, sollte lediglich die Begründung für das Vorgehen aufführen, jedoch sollten inhaltliche Punkte, die zur Anwendung der Methodik notwendig sind, in den jeweiligen Kapiteln benannt werden.)

Das unter Punkt C) auf Seite 19 genannten Vorgehen der Heranziehung von Responderanalysen bedarf des Zusatzes der Relevanzbewertung der Cut-off-Werte, sofern keine validierten bzw. etablierten Kriterien vorhanden sind.

„Für den dritten Schritt der Operationalisierung, der Gesamtaussage zum Ausmaß des Zusatznutzens bei gemeinsamer Betrachtung aller Endpunkte, ist eine strenge Formalisierung nicht möglich, da für die hierzu zu treffenden Werturteile gegenwärtig keine ausreichende Abstraktion bekannt ist.“
Die Gesamtaussage zum Ausmaß des Zusatznutzens ist eine für das Patientenwohl in Deutschland entscheidendes Kriterium, denn in der Abfolge kann ein begründeter Vorschlag für eine Gesamtaussage auf die Entscheidung des G-BA zur Frühen Nutzenbewertung Einfluss nehmen:
Das IQWiG greift unterdessen auf zahlreiche (>30) Nutzenbewertungen von Arzneimitteln gemäß § 35a SGB V zurück, es sollte möglich sein, die Zusammenführung zu einer Gesamtaussage wenigstens in einer allgemeinen Methodik als Rahmen für die Institutsarbeit beschreiben zu können. Inwieweit das Kapitel „3.1.4 Endpunktbezogene Bewertung“ und „3.1.5 Zusammenfassende Bewertung“ hierbei involviert werden können, bleibt offen. Es ist unklar, wie die Zusammenführung konkret stattfindet und eine Abwägung von Nutzen und Risiko im Fokus der verschiedenen Zielgrößen und ggf. Teilpopulationen stattfinden kann.

Um die gesetzliche Vorgabe, die Evidenz-basierte Medizin der Beschluss- bzw. Bewertungspraxis des IQWiG zugrunde zu legen, sicherzustellen, sollten bereits die Rahmenbedingungen für die Arbeitsweise des IQWiG unter Hinzuziehung a) der klinischen Erfahrungen der Ärztinnen und Ärzte b) der Präferenzen und Erfahrungen der Patientinnen und Patienten und c) der verfügbaren Studien/Publikationen transparent erarbeitet werden[3]. Die methodischen Grundlagen mit der Fachöffentlichkeit gemeinsam zu entwickeln ist notwendig. Das Kapitel 3.3.3 bedarf dahingehend einer vollständigen Überarbeitung.

[3] Stallberg, Evidenz-basierte Medizin. PharmR 1/2010

ad 7.3.8 Meta-Analysen

Hilfreich wäre, wenn weitere Literaturstellen eingefügt werden (z.B. Seite 20 erster Abschnitt) und die Literaturstellen auf ihre Aktualität hin überprüft werden (z.B. ist die Literaturstelle [110] nicht aktuell).

Um z. B. den Satz *„Die Spezifizierung, wann eine „zu große" Heterogenität vorliegt, ist kontextabhängig"* transparent zu gestalten, könnte man auf das Kapitel „9.5 Heterogeneity" des aktuellen Cochrane Handbuchs verweisen.
Dieser Satz ist besonders aufgefallen, da ein wichtiger Textabschnitt, nämlich *„…und erfolgt für die jeweiligen Projekte im Berichtsplan"* entfernt wurde. Der BAH empfiehlt, dass dieser entfernte Inhalt wieder aufgeführt wird, da ansonsten die angestrebte Transparenz Schaden nehmen könnte.

Wie bereits in „ad 3.1.4 Endpunktbezogene Bewertung" angemerkt, fällt auf, dass das neu aufgenommene *„Prädiktionsintervall"* starke Betonung/Einbindung erhält. Im Vergleich zu anderen Inhalten erscheint eine Überbetonung stattzufinden. Jedoch stellt sich die Frage, inwieweit das Prädiktionsintervall überhaupt praktische Relevanz z.B. in der Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V hat.

Die zu erwartende Stellungnahme der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (gmds) wird z.B. in der Frage nach dem Stellenwert und der korrekten Einarbeitung des methodischen Ansatzes des Prädiktionsintervalls vom BAH mit hohem Interesse verfolgt werden. Mit Interesse richten der BAH ebenso seinen Blick u.a. auf die Frage, ob die Darstellung des Prädiktionsintervalls in Forest Plots von Meta-Analysen mit zufälligen Effekten bei einer Basis des Vorhandenseins von mindestens 4 Studien sinngemäß ist.

ad „Anhang: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens"

Die Wortwahl „Rationale" erscheint nicht passend (s. ad 3.3.3), da der aktuelle Duden das Wort nicht führt.

Um juristisch exakt zu zitieren und eine exakte Ausarbeitung des IQWiG gewährleisten zu können, sollte in dem zweiten Satz *„Gemäß § 5 Abs. 4 Satz 1 ..."* (S. 26) die Ergänzung *„und mit welcher Wahrscheinlichkeit"* vorgenommen werden.

Ebenso sollte entsprechend § 5 Abs. 7 der AM-NutzenV im dritten Satz (S. 26) unter Punkt (4) korrekt *„Zusatznutzen vorhanden, nicht quantifizierbar"* zitiert werden.

Im letzten Satz der Seite 26 *„Weiterhin erscheint die Zielgröße (gesundheitsbezogene) Lebensqualität, die in § 2 Abs. 3 der AM-NutzenV explizit als Nutzenkriterium formuliert wird, überhaupt nicht in der Kriterienliste für das Ausmaß des Zusatznutzens."*
Diese Darstellung ist fehlerhaft, denn in § 5 Abs. 7 Nr. 1 bis 3 wird die „Lebensqualität" durch Zitierung des § 2 Abs. 3 der AM-NutzenV indirekt eingefügt.

In dem zweiten Abschnitt der Seite 28 wären der Darstellung der Umstrukturierung die weiteren in der Verordnung genannten Zielgrößen, Überlebenszeit und Krankheitsdauer beizufügen, um einer vollständigen exakten Aufzählung zu entsprechen.

Der BAH begrüßt die detaillierte Begründung und Darstellung im Anhang, auch wenn der BAH dieses Vorgehen als nicht ausreichend im Sinne der Evidenz-basierten Medizin beurteilt.

Abschließende Würdigung

Der BAH bedankt sich für die Ausarbeitungen des ersten Teilschritts der Überarbeitung des Methodenpapiers des IQWiG, wobei insbesondere das Kapitel 3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V eine erneute Überarbeitung bedarf. Das methodische Vorgehen, speziell die Herleitung der Schwellenwerte zur Kategorisierung des Ausmaßes eines Effektes und deren Anwendung sowie die Abwägung und Gewichtung von Nutzen und Risiko muss diskutiert werden und von unabhängigen Fachgesellschaften (u.a. der Medizin, der Biometrie, Epidemiologie, Ethik und der Patientenvertreter) bzw. Sachverständigen begutachtet werden.

Im Sinne des gesetzlichen Auftrages, die Bewertung des medizinischen Nutzens nach den international anerkannten Standards der Evidenz-basierten Medizin durchzuführen (gemäß § 139a Abs. 4 SGB V), sollte bereits die Entwicklung der Methodik unter Einbeziehung von Patientinnen und Patienten, Ärztinnen und Ärzten sowie nationaler und internationaler Fachexpertinnen und Fachexperten stattfinden und transparent erarbeitet werden.

Der BAH bittet das IQWiG, seine Stellungnahme bei der weiteren Bearbeitung zu berücksichtigen und steht jederzeit gerne für Rückfragen zur Verfügung. Ebenso wird der BAH sich gerne bei einer mündlichen Anhörung beteiligen.

Bonn, den 22.5.2013
gez. Prof. Dr. Eva Münster, MPH

Bundesverband der Arzneimittel-Hersteller e.V.
Postfach 20 12 55
53142 Bonn
Tel.: +49-228-95745-0
Fax.: +49-228-95745-90
E-Mail: bah@bah-bonn.de
www.bah-bonn.de

A.1.7 – Bundesverband der Pharmazeutischen Industrie e. V. (BPI)

BPI Bundesverband der Pharmazeutischen Industrie e.V.

Friedrichstraße 148
10117 Berlin
www.bpi.de

Telefon: [redacted]
Telefax: [redacted]
E-Mail: [redacted]

Stellungnahme

Stellungnahme des

Bundesverbandes der Pharmazeutischen Industrie zur

Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 des IQWiG sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden 4.1 des IQWiG

in der Entwurfsfassung

vom 18.04.2013

Allgemeine Stellungnahme:

1. **Ganze Stellungnahme zur Anhörung stellen:** Es ist aus Sicht des Verbandes nicht sachgerecht, nur für einzelne Teile der Überarbeitung des Methodenpapiers ein Stellungnahmeverfahren zu eröffnen. Auch wenn es sinnvoll ist, die Methodik schrittweise zu überarbeiten, sollte es möglich sein, alle Kapitel in Kontext zu kommentieren. Die Methodik ist als zusammenhängendes Werk zu sehen, bei den verschiedenen Abschnitten besteht ein innerer Zusammenhang, der in dem nun erfolgenden gestaffelten Stellungnahmeverfahren nicht überblickt werden kann.

2. **Transparenz der Änderungen:** Zur Erleichterung der Stellungnahme sollten alte und neue Passagen gegenübergestellt werden bzw. die Änderungen in geeigneter Weise kenntlich gemacht werden.

3. **Längere Anhörungsfristen:** Die Anhörungsfrist ist mit vier Wochen für eine adäquate Bewertung der Methodik zu kurz. Es sollte im Interesse des Instituts sein, dass sich die Stellungnehmenden hinreichend mit den Änderungen auseinandersetzen können. Sowohl bei den wissenschaftlichen Fachgesellschaften als auch bei den Verbänden sind interne Abstimmungsprozesse einzuhalten, die in dieser kurzen Frist nicht durchlaufen werden können. Es besteht im Hinblick auf das Methodenpapier – möglicherweise im Gegensatz zu konkreten Aufträgen des IQWiG – auch keine Notwendigkeit, für ein Stellungnahmeverfahren eine so kurze Frist zu setzen, die im konkreten Fall durch die dem Institut bekannte Feiertags- und Urlaubssituation im Monat Mai zusätzlich verschärft wird.

4. Bisher kontrolliert das IQWiG beinahe ausschließlich den Fehler, dass einem Medikament ohne Zusatznutzen fälschlicherweise ein Zusatznutzen zuerkannt wird. Hier werden die Anforderungen sehr hoch gesetzt. Dabei wird nicht beachtet, dass durch die hohen Anforderungen die Gefahr groß ist, dass Medikamente mit Zusatznutzen fälschlicherweise im Bewertungsverfahren keinen oder einen zu geringen Zusatznutzen zugestanden wird. Dies ist vor allem vor dem Hintergrund zu sehen, dass die früheren Studien ohne ein Wissen um die heutige Gesetzeslage (AMNOG) durchgeführt wurden. Hier ist eine offene Diskussion mit allen betroffenen Akteuren (Patienten, wissenschaftliche Fachgesellschaften, pharmazeutische Industrie, Politik, etc.) notwendig. Detaillierte Kommentare zu dieser Problematik finden sich auch in den kapitelspezifischen Anmerkungen 3.1.4.-3. und 3.3.3.-1.

5. **Nutzenbewertung von Diagnostika**: Für die Bewertung von Arzneimitteln zu diagnostischen Zwecken im Rahmen der frühen Nutzenbewertung sind im vorliegenden Methodenentwurf keine spezifischen Anforderungen genannt, was dazu führt, dass Diagnostika nach denselben Kriterien bewertet werden wie therapeutische Arzneimittel. Für Diagnostika allein beispielsweise eine Verbesserung hinsichtlich der

BPI Bundesverband der Pharmazeutischen Industrie e.V.

Friedrichstraße 148
10117 Berlin
www.bpi.de

Stellungnahme

Mortalität oder der Morbidität zu zeigen, ist in der Regel nicht möglich, da dies maßgeblich von der nachgeschalteten Therapie abhängig ist. Auch die ZVT wird diese Daten nicht oder nur in Bezug auf eine nachgeschaltete Therapie zeigen können. Hier sind die unterschiedlichen Endpunkte bei den Studien von Diagnostika und Therapeutika ausschlaggebend.
So ist beispielsweise im Hinblick auf ein Organversagen bzw. eine Transplantation die Beurteilung eines Diagnostikums, einschließlich der resultierenden Überlebensrate, welches die Organfunktion misst und somit die Dringlichkeit einer Transplantation anzeigen könnte, abhängig davon ab, ob und wie transplantiert wird.

Es ist daher aus Sicht des Verbandes erforderlich, adäquate methodische Vorgaben zur Bewertung von Diagnostika zu entwickeln. Hierbei ist zu berücksichtigen, dass die erstmalige Eröffnung der Möglichkeit einer Diagnose oder eine Verbesserung einer bestehenden diagnostischen Option für sich genommen – und auch ohne das Vorhandensein einer anschließenden Therapieoption – der maßgebliche Aspekte für die Bewertung des Zusatznutzens eines Diagnostikums sein muss. Patienten haben einen Anspruch darauf, über eine bestehende Erkrankung auch dann informiert zu werden, wenn eine Therapie möglicherweise nicht zur Verfügung steht. Es ist daher nicht adäquat, Diagnostika hinsichtlich des Zusatznutzens vorwiegend im Hinblick auf die mit ihnen verbundene Therapieoption hin zu bewerten.

Besondere Stellungnahme:

Zu 2.1.1 Bericht

A) Ablauf der Berichterstellung
Es ist unklar, warum der Review des Vorberichts zur Qualitätssicherung von einem oder mehreren externen Reviewern mit ausgewiesener methodischer und / oder fachlicher Kompetenz nur noch optional durchgeführt werden soll.

In der Abbildung 1 erscheint die "Hinzuziehung einzelner Patienten / Patientenvertreter" nicht repräsentativ. Vorschlag: "Hinzuziehung relevanter Patientenorganisationen und / oder Fachgesellschaften."

B) Allgemeine Anmerkungen zum Stellungnahmeverfahren (Anhörung)
Es ist die Bekräftigung zu begrüßen, dass die Stellungnahmen in die Entscheidung einzubeziehen sind.

Zu 3.1.4 Endpunktbezogene Bewertung

1. Hier wird statt von einer Ergebnisunsicherheit nur noch von einer Ergebnissicherheit gesprochen. Ersteres wäre jedoch zutreffender, da es immer um Unsicherheiten gehen wird und eine absolute *Sicherheit* nicht erreichbar ist.

2. Um zu beurteilen, ob Effekte gleichgerichtet sind, wurde das Prädiktionsintervall eingeführt. Hierdurch haben sich die Anforderungen an die Aussagesicherheit erhöht. Im Beispiel des folgendem Settings
   - vier Studien mit hoher Ergebnissicherheit,
   - eine Zusammenfassung der Studien in einer Meta-Analyse ist nicht möglich,
   - die Effektrichtung des Großteils der Studien ist dieselbe,
   - die Effekte dieser Studien sind mehrheitlich statistisch signifikant,
   - das Prädiktionsintervall überdeckt den Nulleffekt

   liegt nach der neuesten Änderung bspw. nur noch ein Hinweis vor. Nach der bisherigen Methodik hat dies zu einem Beleg geführt. Worin ist diese Verschärfung begründet?

3. **Berücksichtigung auch des β-Fehlers (Fehler zweiter Art):** In der Statistik sind zwei grundsätzliche Fehler zu unterscheiden, die bei der Testung einer Hypothese möglich

sind. Man spricht hier vom α- und β-Fehler. Auf die frühe Nutzenbewertung bezogen, heißt dies:

a) *Einem Arzneimittel ohne Zusatznutzen wird fälschlicherweise ein Zusatznutzen zugesprochen (α-Fehler).*
Wird der α-Fehler, wie hier, auf die gesamte Methodik zur Ermittlung des Zusatznutzens von Arzneimitteln bezogen, so haben natürlich auch die Anforderungen an Surrogatparameter, die Akzeptanz von post hoc Subgruppen und weitere methodische Detailfragen einen Einfluss darauf, ob im Rahmen der gesamten frühen Nutzenbewertung ein α-Fehler begangen wird oder nicht.

b) *Einem Arzneimittel wird das Fehlen eines Zusatznutzens attestiert, obwohl tatsächlich ein Zusatznutzen vorhanden ist (β-Fehler).*
Auch der β-Fehler bezieht sich nach unserer Interpretation hier nicht allein auf die statistischen Testverfahren, sondern beinhaltet alle weiteren methodischen Detailfragen, die Rückwirkungen auf die Sicherheit des Ergebnisses haben.

Gerade im Interesse der Patienten ist ein rationaler Umgang mit der unvermeidbaren Unsicherheit erforderlich, damit den Menschen im Krankheitsfall – wie vom Gesetzgeber gewünscht – die besten und wirksamsten Arzneimittel zur Verfügung stehen. Ziel sollte es deshalb gerade auch im Sinne der Patienten und der Gesellschaft sein, nicht eine maximale oder minimale Ergebnissicherheit, sondern eine optimale Ergebnissicherheit anzustreben.

Einen ersten Anhaltspunkt wo diese liegt, gibt die vom IQWiG gewählte Beschreibung der Evidenzbasierten Medizin (EbM) als Strategie, „[…] *die vor Fehlentscheidungen und falschen Erwartungen schützen* […]“ soll. Da sich Fehlentscheidungen nicht vollständig vermeiden lassen, weil, wie gezeigt, ein Trade-off-Effekt zwischen den beiden prinzipiellen Fehlermöglichkeiten besteht, ist es ein logisches Ziel, die Wahrscheinlichkeit einen α- oder β-Fehler zu begehen, zu minimieren.

Die Minimierung der Wahrscheinlichkeit einen α- oder β-Fehler zu begehen reicht aus einer ökonomischen Perspektive jedoch noch nicht aus, um die optimale Ergebnissicherheit zu definieren. Die AMNOG-Ziele sollten nicht nur quantitativ, sondern vor allem qualitativ erfüllt werden. Dies macht es notwendig, zusätzlich zu untersuchen, mit welchen volkswirtschaftlichen Kosten das Begehen eines der beiden prinzipiellen Fehler behaftet ist.

4. Das Methodenpapier des IQWiG legt ferner fest, dass die Belegbarkeit des Zusatznutzens oder –schadens in den 4 Kategorien „Beleg“, „Hinweis“, „Anhaltspunkt“ oder „keines davon“ erfolgt.

Grundsätzlich muss festgestellt werden, dass diese Kategorisierung weder hergeleitet und nachvollziehbar begründet wird, noch international angewendeten HTA-Standards entspricht (wie es der Gesetzgeber in § 139a Abs. 4 SGB V explizit fordert).

Auch die vom IQWiG verwendete Kategorisierung zum Ausmaß der qualitativen Ergebnissicherheit ist nicht eindeutig nachvollziehbar, da es keine Operationalisierung z.B. zur Unterscheidung von hohem oder niedrigem Verzerrungspotential gibt und die Gesamteinschätzung dadurch nicht immer nachvollziehbar und transparent ist - und unter Umständen sogar subjektiv gefärbt ist.

Zur Ableitung eines „Belegs für Zusatznutzen“ fordert das IQWiG mindestens 2 Studien mit statistisch signifikantem Ergebnis der Metaanalyse bei vorhandener Homogenität. Zur Ableitung der Beleglage bei heterogenen Ergebnissen aus Meta-Analysen fordert das IQWiG gleichgerichtete Effekte in Studien mit den Bedingungen: das Gesamtgewicht dieser „gleichgerichteten“ Studien soll ≥ 80% sein und mindestens 2 dieser Studien sollen statistisch signifikante Ergebnisse zeigen. An dieser Stelle fehlt eine Begründung für die Festlegung des Gesamtgewichtes auf mindestens 80% für die Feststellung gleichgerichteter Effekte.

Sofern es sich nicht um primäre Endpunkte der betrachteten Studien handelt, war die Fallzahlplanung der Studien nicht auf das Erreichen statistischer Signifikanz ausgelegt. In dieser Situation ist Signifikanz, wenn überhaupt, nur in einer Meta-Analyse nachweisbar und die Forderung, dass zwei Studien einzeln signifikant sein müssen, nicht zu begründen.

Für den Fall von weniger Studien oder schwächerer Ergebnissicherheit dieser Studien behält sich das IQWiG die Kategorien „Hinweis“ und „Anhaltspunkt“ vor. Bei einer einzelnen Studie mit hoher qualitativer Ergebnissicherheit wird dementsprechend kein „Beleg“ mehr, sondern nur noch ein „Hinweis“ zuerkannt. Das IQWiG stellt in Ausnahmefällen die Möglichkeit in Aussicht, auch bei Vorliegen einer einzelnen Studie einen „Beleg“ zu erhalten, wenn bestimmte Voraussetzungen erfüllt sind, die in den „Points to consider on application with 1. meta-analyses; 2. one pivotal study“ der European Medicines Agency (EMA), der europäischen zentralen Zulassungsbehörde, näher ausgeführt sind. Dies ist sehr zu begrüßen, da gerade für Patienten, für die sonst keine entsprechenden Therapiealternativen mehr vorhanden sind, aus ethischen Gründen auch keine neuerlichen randomisierten Studien mit einem Medikament möglich sind, das seine Überlegenheit bereits unter Beweis gestellt hat. Dies ist bspw. häufig der Fall bei onkologischen Erkrankungen, wo es unethisch wäre, das einzige noch wirkende Produkt den Patienten im Kontrollarme vorzuenthalten. Schon aus diesem Grund würde eine entsprechende zweite Studie durch Ethikkommissionen nicht genehmigt werden, weshalb die EMA in diesen Fällen bereits bei Vorliegen von nur einer Phase III Studie in Anwendung des oben zitierten Dokuments eine Zulassung erteilt.

In der Vergangenheit musste jedoch leider festgestellt werden, dass das IQWiG an dieser Stelle nicht immer seiner eigenen Methodik folgt.

Zusätzlich sollte berücksichtigt werden, dass der Anteil der in eine Studie eingeschlossenen Patienten, bezogen auf die Gesamtpopulation der an einer bestimmten Krankheit leidenden Patienten, in einer einzelnen Studie bei einer selteneren Erkrankung oftmals höher ist als bei einer anderen, häufiger auftretenden Indikation die Patientenanteile aller Phase III-Zulassungsstudien gemeinsam. Eine zu starre Fokussierung auf die bloße Anzahl der Studien, die zur Zulassung geführt haben, ohne Berücksichtigung der Repräsentativität, benachteiligt somit unter Umständen Patienten, die an einer selteneren Erkrankung leiden.

In Tabelle 2, S. 11, wird die Beleglage für unterschiedliche Aussagesicherheiten nochmals tabellarisch dargestellt. In vorangehenden Diskussionen mit dem IQWIG wurde die Frage aufgeworfen, ob bei einer einzelnen Studie mit direktem Vergleich gegenüber der zweckmäßigen Vergleichstherapie die Beleglage durch die Hinzunahme indirekter Vergleiche aufgewertet werden kann, um damit eine höhere Stufe der Aussagesicherheit zu erreichen. Wir schlagen daher vor, Tabelle 2 um Szenarien zu ergänzen, in denen die Aufwertung der Aussagesicherheit unter zusätzlicher Berücksichtigung indirekter Vergleiche abgebildet ist.

Indirekte Vergleiche können darüber hinaus, sofern sie mit direkten Vergleichen konsistent sind, auch hilfreich sein, die Präzision der Schätzung von Abständen zur zweckmäßigen Vergleichstherapie zu verbessern.

Im Folgenden wird auf Seite 12 präzisiert, dass „die für einen Beleg notwendige Bestätigung (…) eines statistisch signifikanten Ergebnisses einer Studie hoher qualitativer Ergebnissicherheit (…) durch (…) Ergebnisse mäßiger (…) qualitativer Ergebnissicherheit erbracht werden“ können. An dieser Stelle fehlt eine entsprechende Rationale dafür, warum der Bereich für das Gewicht einer Studie hoher qualitativer Ergebnissicherheit auf 25% bis 75% festgelegt wird.

5. **Differenzierung der Bewertungsmethodik:** IQWiG und G-BA verfolgen einen methodischen Ansatz, nach dem die gleichen Anforderungen für jedes zu bewertende Arzneimittel weitgehend unterschiedslos gelten sollen. Zudem wird die Ergebnisunsicherheit im Wesentlichen nur beschrieben (Anhaltspunkt, Hinweis, Beleg) ohne daraus direkte Folgen abzuleiten.
   Es ist aus Sicht des Verbandes erforderlich, die Methodik 4.0 (und weiterer Methodenpapiere) des IQWiG entsprechend weiter zu entwickeln. Welche Ergebnissicherheit der Studienlage vom IQWiG bspw. gerade noch als Anhaltspunkt für einen Zusatznutzen anerkannt wird, sollte nicht fix sein, sondern u. a. von folgenden Parametern abhängen:

a. der Schwere der Erkrankung (Relevanz für den Patienten),
b. der Anzahl und der Güte der therapeutischen Alternativen und
c. der wahrscheinlichen Effektgröße (auch unter Berücksichtigung irreversibler, schwerer,
nicht behandelbarer Nebenwirkungen).

Da es nicht möglich ist, für jede Erkrankung individuelle Anforderungen festzulegen, könnten beispielsweise interventions- oder indikationsspezifische „Cluster“ (Gruppen) gebildet werden.

3.1.5 Zusammenfassende Bewertung

1. **Rechtliche Grundlage für Saldierung von Nutzen und Nebenwirkungsrisiken (in der Terminologie des IQWiG als „Schaden“ bezeichnet) nicht ersichtlich:** Nach § 35 a Abs. 1 Satz 2 SGB V ist das Zusatznutzen gegenüber der zweckmäßigen Vergleichstherapie, das Ausmaß des Zusatznutzens und seine therapeutische Bedeutung zu bewerten. Zwar enthält das SGB V keine Legaldefinition des Zusatznutzens, aber § 35 a Abs. 1 Satz 7 Nr. 2 SGB V ermächtigt das BMG, in einer Rechtsverordnung die Grundsätze für die Bestimmung des Zusatznutzens festzulegen. Von dieser Ermächtigung hat das BMG Gebrauch gemacht. § 2 Abs. 3 AMG-NutzenV definiert den Nutzen als für den Patienten relevanten therapeutischen Effekt, insbesondere hinsichtlich der Verbesserung des Gesundheitszustandes, der Verkürzung der Krankheitsdauer, der Verlängerung des Überlebens, der Verringerung von Nebenwirkungen oder einer Verbesserung des Lebensqualität. Der Zusatznutzen wird nach § 2 Abs. 4 AM-NutzenV definiert als ein Nutzen i. S. d. Abs. 3, der quantitativ oder qualitativ höher ist als ein Nutzen, den die zweckmäßige Vergleichstherapie aufweist. Eine Saldierung des festgestellten Zusatznutzens mit dem Nebenwirkungsrisiko eines Arzneimittels findet daher in diesen Legaldefinitionen keine Grundlage.

   Vielmehr sieht § 5 Abs. 5 Satz 1 AM-NutzenV vor, dass der Zusatznutzen festgestellt wird als Verbesserung der Beeinflussung patientenrelevanter Endpunkte zum Nutzen gem. § 2 Abs. 3 der AM-NutzenV. Dass hier Nutzen und Risiken des Arzneimittels im angemessenen Verhältnis zueinander stehen, ergibt sich zwingend aus der arzneimittelrechtlichen Zulassung, weil bei einem ungünstigen Nutzen-Risiko-Verhältnis keine arzneimittelrechtliche Zulassung erteilt worden wäre (vgl. für die deutsche Rechtslage § 25 Abs. 2 Nr. 5 AMG). Diese Bewertung der Zulassungsbehörde ist auch für die sozialrechtliche Nutzenbewertung von G-BA und IQWIG vorgreiflich (vgl. §§ 5 Abs. 3 Satz 2, 7 Abs. 2 Satz 6 AM-NutzenV).

2. **Vorgehen zur Saldierung von Nutzen und Nebenwirkungsrisiken (in der Terminologie des IQWiG als „Schaden“ bezeichnet) weiterhin unklar:** Dieses Kapitel behandelt eher theoretisch, welche Möglichkeiten das IQWiG sieht, die verschiedenen

Nutzen- und Schadensaspekte innerhalb eines Maßes zu aggregieren, z. B. durch die Gewichtung patientenrelevanter Endpunkte zur Bildung eines Summenscores. Eine detailliertere Ausführung fehlt jedoch. Es wird betont, dass die Methodik der Aggregation möglichst in einem Berichtsplan oder Vorbericht spezifiziert werden sollte. Da der Prozess der frühen Nutzenbewertung gem. § 35a SGB V solch einen Berichtsplan bzw. Vorbericht nicht vorsieht, ist weiterhin unklar, wie in den Verfahren zur frühen Nutzenbewertung vorgegangen wird. Im Sinne einer Planungssicherheit für den pharmazeutischen Unternehmer wäre eine konkretere Beschreibung des Vorgehens wünschenswert. Grundsätzlich stellt sich jedoch auch hier die Frage nach der ethisch-moralischen Legitimation des IQWiG, solche Werteentscheidungen zu treffen. Bislang wird nämlich in dieser Diskussion außer Acht gelassen, wo die Präferenzen des betroffenen Patienten in einer bestimmten Situation liegen. Im Falle einer schwerwiegenden, zum Tode führenden Krankheit, können in der Wahrnehmung des Patienten Nebenwirkungen durch eine bestimmte Verlängerung des Lebens möglicherweise aufgewogen werden. Solche Präferenzen können von Erkrankung zu Erkrankung je nach Schweregrad der Erkrankung und der möglichen Nebenwirkungen selbst sehr stark variieren, aber auch innerhalb einer Erkrankung können sich die individuellen Wahrnehmungen und Entscheidungen sehr unterscheiden.

3. **Präferenzbasierte Instrumente wichtige Tools für Berücksichtigung von Patientenpräferenzen:** Es ist hinlänglich belegt, dass die Berücksichtigung der Patientenpräferenzen einen wesentlichen Teilaspekt der objektiven Bewertung der gesundheitsbezogenen Lebensqualität darstellt. Dabei sind präferenzbasierte Instrumente den generische Lebensqualitätsinstrumente zu bevorzugen da nur dort der direkte Bezug zu indikations- und/bzw. interventionsbezogenen Eigenschaften hergestellt werden kann (Bridges et al. 2007).[1]

   Basis des klinischen und des ökonomischen Entscheidungskalküls ist der Nutzen für die Patienten. Der Nutzen für Zulassung, Erstattungsfähigkeit und Preisfestsetzung wird jedoch oft in zwei unterschiedlichen Studienwelten dokumentiert, d.h. je nach Wissenschaftsdisziplin und Anwendungskontext werden der Nutzenbewertung unterschiedliche Kriterien zugrunde gelegt.

   Klinische Perspektive: Die Evidenzbasierte Medizin fokussiert auf die Messung und Interpretation patienten-relevanter Endpunkte, um den Erfolg alternativer Strategien zu vergleichen. Patienten-relevante Endpunkte sind messbare klinische Erfolgskriterien der Morbidität, Mortalität und Lebensqualität (§§ 35a, 35b SGB V). Daneben können der interventionsbezogene Aufwand und die Patientenzufriedenheit als weitere Erfolgsparameter in die Bewertung einbezogen werden. Multikritielle Bewertungsverfahren ermöglichen es diese beiden Studienwelten zu kombinieren.

[1] Bridges, J., Onukwugha, E., Johnson, F. R., & Hauber, A. B. (2007). Patient preference methods—a patient centered evaluation paradigm. ISPOR connections, 13(6), 4-7.

In den Allgemeinen Methoden 4.0 wird explizit der Einbezug der Methoden der multikriteriellen Entscheidungsfindung oder der Präferenzerhebung in den Kapiteln 3.1.1 sowie vertiefend in 3.1.4 bei für die Definition und Hierarchisierung von patientenrelevanten Endpunkten beschrieben. Bei der Aktualisierung der Allgemeinen Methoden wird nur das Kapitel 3.1.4 aufgegliedert und die beiden Methoden (Analytic Hierarchy Process (AHP) und die Conjoint-Analyse (CA)) sind nunmehr im Kapitel 3.1.5 zu finden. Das Kapitel 3.1.1 bleibt unverändert. Sollte die ausführliche Darstellung der beiden Methoden (AHP sowie CA) im Rahmen der IQWiG-Piloten veröffentlicht werden, kann an dieser Stelle auf die ausführliche Beschreibung verzichtet werden.

Da für die Ableitung der Gesamtbewertung der Beleglage, Nutzen und Schaden, wie in 3.1.5 gefordert, einzeln dargestellt bzw. sogar gegeneinander abgewogen werden sollen, sind multikriterelle Erhebungsinstrumente wie dargestellt anzuwenden. Sie ermöglichen es, den Patientennutzen auf Basis von Abwägungsentscheidungen (sogenannte Trade-offs) von patientenrelevanten Endpunkten in Form einer Nutzen-Schadensabwägung darzustellen. Diese patientenseitige Nutzen-Schadensabwägung kann dabei sowohl Mortalitäts- als auch Morbiditätsaspekte enthalten (je nach Indikation, Intervention und Komparator).

Ferner wird es damit möglich, den nach § 2 Abs. 3 AM-NutzenV maßgeblichen Patienten-Nutzen als *„patientenrelevanten therapeutischen Effekt insbesondere hinsichtlich der Verbesserung des Gesundheitszustandes, der Verkürzung der Krankheitsdauer, der Verlängerung des Überlebens, der Verringerung der Nebenwirkungen oder der Verbesserung der Lebensqualität“* angemessen zu berücksichtigen“

Zu 3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V und Anhang „Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens“

Am 20. Juni 2012 hat auf Einladung des Gemeinsamen Bundesausschusses ein Methodik-Workshop zur Klassifizierung des Zusatznutzens in der frühen Nutzenbewertung stattgefunden. Neben den Vertretern der Herstellerverbände BAH, BPI, Pro Generika und vfa und den Vertretern der Bänke (GKV-Spitzenverband, KBV und DKG) waren Patientenvertreter des G-BA, das IQWiG, die AWMF, die AkdÄ sowie das Paul-Ehrlich-Institut (PEI) und das BMG eingeladen.

Die dort geführten Diskussionen ergaben, dass die vom IQWiG entwickelte Bewertungssystematik im Anhang A des IQWiG-Bewertungsberichts zum Wirkstoff Ticagrelor aus Sicht der vertretenen Institutionen auf einer ungenügenden Studienbasis und einer nicht ausreichenden Literaturrecherche basiert und zu holzschnittartig ist, um indikations- bzw. interventionsspezifisch in unterschiedlichen Versorgungssituationen auch im Hinblick auf Patientenpräferenzen zu adäquaten Bewertungsergebnissen zu kommen.

Zudem wurde deutlich, dass die der Bewertungssystematik innewohnenden Wertentscheidungen nicht eigenständig durch das IQWiG als wissenschaftliches Institut definiert werden können, sondern in einem Prozess zu erarbeiten sind, der u. a. auch eine breitere Diskussion mit betroffenen Patienten und klinisch tätigen Ärzten beinhalten muss. Es wurde anerkannt, dass das IQWiG – auch aus der Situation heraus, dass keine andere Vorgabe zur Erfüllung der gesetzlichen Vorgaben vorhanden war – einen ersten Ansatz als Diskussionsgrundlage entwickelt hat. Dieser Ansatz bedürfe aber einer Weiterentwicklung und einer Validierung. Sollte das IQWiG sich entscheiden, die Bewertungssystematik weiterzuverfolgen, so wäre es zudem unerlässlich, dass diese - nach entsprechendem wissenschaftlichen Diskurs - Eingang in sein Methodenpapier findet.

Der Anhang des vorliegenden IQWiG-Papiers lehnt sich stark an den Anhang A zur Dossierbewertung im Rahmen der Nutzenbewertung des Wirkstoffs Ticagrelor [2] an. In diesem Teil der Bewertung wurde die quantitative Operationalisierung der Vorgaben der AM-NutzenV festgelegt und teilweise hergeleitet.

Bei näherer Betrachtung des zur Anhörung gestellten Anhangs sind keine maßgeblichen Veränderungen im Hinblick auf die zum damaligen Zeitpunkt diskutierte Methodik im Anhang des Bewertungsberichts zum Ticagrelor-Dossier zu erkennen.

Dies ist umso bedauerlicher, als auf dem Methodik-Workshop konkrete Änderungsvorschläge gemacht worden sind. Es ist nicht zu erkennen, dass sich das IQWiG mit diesen Vorschlägen im Vorfeld der Veröffentlichung des vorliegenden Entwurfs auseinandergesetzt hat. In jedem Fall sind keine maßgeblichen Veränderungen gegenüber dem damaligen Sachstand erkennbar.

Der BPI möchte daher die bereits zum damaligen Zeitpunkt vorgebrachten Kritikpunkte erneuern und, wo erforderlich, ergänzen:

a. Gesetzliche Vorgaben beachten!

Das IQWiG führt im vorliegenden Entwurf des Anhangs aus, dass die in § 5 Abs. 7 der AM-NutzenV vorhandenen Kriterien zum Teil eindeutig und zum Teil weniger eindeutig bestimmt seien. Weiterhin folgert es, dass in einigen Fällen Kategorien nicht für alle aufgeführten Kriterien erschöpfend besetzt seien. Es erscheine daher geboten und sinnvoll, die gesetzlich vorgegebene Kriterienliste eigenständig anzupassen und zu vervollständigen.

In Bezug auf die Zielkriterium „Überlebensdauer“ interpretiert das IQWiG, dass der Verordnungsgeber in der AM-NutzenV im Falle einer erheblichen Verlängerung einen erheblichen und im Falle einer moderaten Verlängerung einen beträchtlichen Zusatznutzen vorgesehen habe und folgert daraus, dass eine Zuordnung zur Nutzenkategorie gering „vergessen“ worden sei. Diese Sichtweise teilt der BPI nicht. Eine eigenständige

[2] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Ticagrelor: Nutzenbewertung gemäß §35a SGB V; Dossierbewertung; Auftrag 11-02

„Ergänzung“ von Zusatznutzenkategorien über das gesetzlich vorgesehene Maß hinaus steht dem IQWiG nicht zu und muss daher unterbleiben.

In Bezug auf die Aspekte „Heilung“ und „spürbare Linderung der Erkrankung“ schlägt das IQWIG diese im ersten Fall der Zielgröße „Überlebensdauer“ (Mortalität) und im zweiten Fall der Zielgröße „schwerwiegende Symptome“ (Morbidität) zu. Dies ist aus Sicht des BPI im Hinblick auf die gesetzlichen Vorgaben unzulässig. Im Falle einer Heilung ist gemäß der AM-NutzenV in jedem Fall von einem erheblichen Zusatznutzen auszugehen, der Verordnungsgeber nimmt für dieses Kriterium im Hinblick auf die Nutzenkategorien „beträchtlich“ und „gering“ keine weitere Differenzierung vor. Der Begriff taucht in der AM-NutzenV bei der Unterlegung dieser Nutzenkategorien mit Zielkriterien nicht auf. Das Vorgehen des IQWiG würde dazu führen, dass die Heilung einer Erkrankung über die Zuordnung zur Zielgröße „Überlebensdauer“ einer differenzierenden Bewertung zugänglich gemacht werden würde, den die AM-NutzenV für diesen Parameter nicht vorsieht. Zudem sprechen auch inhaltliche Gründe gegen die Einbeziehung des eigenständigen Kriteriums „Heilung“ in das Kriterium „Überlebensdauer“.

Insbesondere im Fall von Arzneimitteln, die zur Akuttherapie einer nicht-tödlichen Erkrankung eingesetzt werden, kann bei einer kurzfristigen Einnahme eine Heilung einer Erkrankung erreicht werden, ohne dass hinsichtlich der Verlängerung der Überlebensdauer Effekte entstehen. Als Beispiel sei ein neues Antibiotikum genannt, das kurzfristig eingesetzt zu einer Heilung einer ansonsten nicht-tödlichen, aber schwerwiegenden Erkrankung führt. Würde für dieses Präparat das eigenständige Kriterium „Heilung“ entfallen, müsste der Zusatznutzen über die Verlängerung der Überlebensdauer nachgewiesen werden, was im betreffenden Fall nicht durch Studien darstellbar sein dürfte. Auch im Fall der „spürbaren Linderung einer Erkrankung“ ist dieses Kriterium eigenständig zu erhalten.

Bezogen auf das Kriterium „Lebensqualität“ teilt der BPI die Auffassung des IQWiG, dass hier eine Ergänzung vorzunehmen ist. Dies ist offensichtlich, da die Zielgröße in § 3 der AM-NutzenV erwähnt wird, bezogen auf die Nutzenkategorien in § 5 Abs. 7 der AM-NutzenV dann aber vollständig fehlt. Die Untergliederung der Zielgröße und der namentlichen Beschreibung hinsichtlich der Zuordnung zu den Nutzenkategorien „erheblich“ und „beträchtlich“ kann gefolgt werden. Ein geringer Zusatznutzen ist für „jegliche Verbesserung der Lebensqualität“ anzuerkennen.

b. Die Festlegung der Effektstärken zum Erreichen von Ausmaßkategorien ist weiterhin subjektiv.

Das Vorgehen zur Operationalisierung des Ausmaßes des Zusatznutzens wurde vom IQWiG nicht näher erläutert. Die Grundlage für die verwendeten Schwellenwerte stellen weiterhin die Ausführungen im Anhang A des Ticagrelor-Berichts dar (S. 86-88), in denen die Effektstärken zum Erreichen bestimmter Ausmaßkategorien für die Endpunkte Mortalität, Schweres Unerwünschtes Ereignis und Unerwünschtes Ereignis postuliert werden. Eine systematische Erfassung möglicher vergleichbarer Operationalisierungen im internationalen Kontext und eine Evaluation der damit verbundenen Erfahrungen haben nicht stattgefunden.

Der verwendeten Literatur, auf der die Operationalisierung beruht, liegt keine systematische Literaturrecherche zugrunde. Als Quelle für die verwendete Effektstärke 50% („Erheblicher Zusatznutzen, Mortalität") wird Djulbegovic et al. (2008)[3] angeführt, der Behandlungen mit einer RR von <=50% als „Breakthrough interventions" bezeichnete. Somit basiert das relative Risiko für eine Verbesserung der Mortalität in der Nutzenkategorie „erheblich" – und damit der „Anker" für die horizontalen und vertikalen Differenzierungen hinsichtlich der relativen Risiken in der Tabelle NT6 – auf Djulbegovic et al. als einziger (!) Literaturquelle, die auf einer sehr spezifischen Selektion von onkologischen Studien aus den Jahren 1955-2006 beruht. Ferner ist bei der qualitativen Würdigung dieser Quelle kritisch anzumerken, dass die Autoren der Studie selbst relativierend von einem gedanklichen Konstrukt und Vorschlag sprechen, die der weiteren wissenschaftlichen Validierung und Diskussion bedürfen.

Es ist nicht angebracht, aufgrund einer einzigen Publikation weitreichende Kriterien für die Beurteilung des Ausmaßes des Zusatznutzens abzuleiten. Dieses Vorgehen widerspricht den internationalen Standards der evidenzbasierten Medizin.

Das IQWiG hat mit der Festlegung eines relativen Risikos von 0,50 als Wert für die Feststellung einer „erheblichen Verlängerung der Überlebensdauer" einen Anker gelegt. Das IQWiG führt hierzu aus, dass ein relatives Risiko von 0,50 von Djulbegovic et al. als Anforderung für eine „Durchbruchsinnovation" postuliert wird. Es ist nicht ersichtlich, warum das nicht weiter bestimmte Kriterium der „Durchbruchsinnovation" für die Festlegung eines erheblichen Zusatznutzens heranzuziehen wäre. Dieser Ausdruck kommt in der AM-NutzenV nicht vor und eignet sich daher nicht für eine Operationalisierung des Ausmaßes des Zusatznutzens. Dort heißt es vielmehr wörtlich: „*Ein erheblicher Zusatznutzen liegt vor, wenn eine nachhaltige und gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte große Verbesserung des therapierelevanten Nutzens im Sinne von § 2 Absatz 3 erreicht wird, insbesondere eine Heilung der Erkrankung, eine erhebliche Verlängerung der Überlebensdauer, eine langfristige Freiheit von schwerwiegenden Symptomen oder die weitgehende Vermeidung schwerwiegender Nebenwirkungen.*"

Die Operationalisierung ist aus den Definitionen des Gesetzes abzuleiten. Es ist in jedem Fall eindeutig, das unter Berücksichtigung der zitierten Definition eines erheblichen Zusatznutzens in der AM-NutzenV eine unzulässige Verengung des Willens des Verordnungsgebers erfolgen würde, wenn diese Nutzenkategorie nur von einer Durchbruchsinnovation erreicht werden könnte.

Die postulierten Risikoabsenkungen wurden mit Hilfe einer vereinfachten Formel (zitiert wird als Quelle Fleiss et.al., die tatsächliche Quelle war das SAS-Handbuch) in die Obergrenze für das 95%-Konfidenzintervall um den RR-Schätzer umgerechnet (Tabelle 32, Ticagrelor-Bericht). Die Obergrenzen der Konfidenzintervalle wurden unverändert in das vorliegende Dokument übernommen (Tabelle NT1).

[3] Djulbegovic B, Kumar A, Soares HP, Hozo I, Bepler G, Clarke M et al. Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute; sponsored cooperative oncology groups, 1955 to 2006. Arch Intern Med 2008; 168(6): 632-642.

Die postulierten Grenzen wurden in der Zwischenzeit nicht weiter objektiviert. Es wurden auch keine weiteren empirischen Belege für die Angemessenheit der Grenzwerte erbracht. Die vom IQWiG zitierte Quelle Djulbegovic et al. gibt an, dass nur 12 von relevanten 614 Studien (=2%) den postulierten Grenzwert von 50% für die RR-Absenkung erreichten. Zusätzlich 116 von 766 insgesamt untersuchten Studien erreichten den Grenzwert nicht, wurden aber von den Forschern als „Breakthrough" bezeichnet (Figure 3). Auf das IQWIG-Klassifikationsschema angewendet bedeutet dies, dass 7 von 8 Studien mit bedeutendem Fortschritt bei der Mortalitätssenkung nicht in die Kategorie „Erhebliche Nutzensteigerung" eingeordnet werden, weil sie das subjektiv festgelegte Effektivitätskriterium nicht erfüllen. Für die übrige Klassifikation muss ebenfalls angenommen werden, dass die Kategorie „erhebliche Nutzensteigerung" nicht oder nur selten erreicht wird, auch wenn aufgrund des p-Wertes der entsprechenden Tests möglicherweise ein hochsignifikantes Ergebnis vorliegt.

c. Die Monte-Carlo-Simulation zur Bestimmung des Zusammenhangs von Effektgröße und Schwellenwert trägt nicht zur Objektivierung bei.

Die im Ticagrelor-Bericht verwendete Formel zur Bestimmung der Obergrenzen der Konfidenzintervalle für die Risk Ratio wurde kritisiert. Vom IQWiG wird im vorliegenden Bericht eingeräumt, dass der Algorithmus, der dieser Formel zugrunde liegt, auf nicht publizierter SAS-Software beruht. Anstelle der Formel treten nun die Ergebnisse einer Monte-Carlo-Studie, deren Ergebnisse in Tabelle NT6 graphisch dargestellt werden. Bei vorgegebener Obergrenze des Konfidenzintervalls und bei vorgegebenem Basisrisiko der Kontrollgruppe wird die erforderliche Risikoabsenkung durch Simulation ermittelt. Insgesamt sind damit die Anforderungen an den Gesamteffekt gelockert worden (z.B. 0,53 bei hohem Basisrisiko, 0,58 bei niedrigem Basisrisiko statt fest 0,5 im Ticagrelor-Bericht). Die subjektiven Grundfestlegungen für die Obergrenzen der Konfidenzintervalle aus dem Ticagrelor-Bericht bleiben aber weiterhin gültig.

Die Monte-Carlo-Analyse selbst ist weitgehend undokumentiert und somit auch nicht replizierbar. Damit ist auch der Anspruch der Wissenschaftlichkeit der Vorgehensweise, den das IQWiG an sich selbst stellt, nicht überprüfbar. Es fehlen Angaben zur Anzahl der Simulationen, zum Simulationsdesign, zur Replizierbarkeit der Zufallszahlen, zur Art des verwendeten Tests (Chi-Quadrat, Fisher's Test, feste oder zufällige Gruppengröße) und zur simulierten Stichprobengröße. Die Ergebnisse stellen nur (graphisch!) Mittelwerte, aber keine Streuungen dar. Die kurzen Angaben auf Seite 35 (alpha=0,05, Power=90%) sind in jedem Fall unzureichend.

Die IQWiG wird aufgefordert, die Berechnung der Schwellenwerte nachvollziehbar darzustellen. Inhalt und Methodik der Monte-Carlo-Simulation sind genauer zu beschreiben.

d. Die berechneten Schwellenwert-Bereiche implizieren über die Power-Annahme (90%) Vorgaben für die Versuchsplanung, die nicht eindeutig dokumentiert wurden.

Während noch im Ticagrelor-Bericht die Power-Annahme bei der Berechnung der Schwellenwerte nicht genannt wurde (S. 89), wird im vorliegenden Bericht bei der Fallzahlberechnung der simulierten Stichproben von einer Power von 90% ausgegangen. Wissenschaftliche standardisierte Vorgaben für die Festsetzung der Power bestehen nicht – die Power wird studienspezifisch festgelegt. Allgemein wird eine Powervorgabe von mindestens 80% bei der Studienplanung akzeptiert[4].

Ergebnis dieser Art der Bestimmung der Grenzwerte ist die Tatsache, dass Studien, deren Stichprobengröße mit einer Power von 80% geplant wurden und die erforderliche Risikoabsenkung erreichen, trotzdem in eine niedrigere Nutzenkategorie eingestuft werden. Der Grund dafür sind die breiteren Konfidenzintervalle aufgrund der niedrigeren Fallzahl.

Das IQWiG wird aufgefordert, die Powerannahme bei der Berechnung von notwendigen Effektstärken an die akzeptierte wissenschaftliche Praxis bei der Stichprobenplanung anzupassen bzw. zu begründen, warum es von der üblichen Vorgabe von 80 % abweicht.

e. Einbettung in einen wissenschaftlichen Diskussionsprozess

Des Weiteren ist nicht nachvollziehbar dargestellt, auf welcher Basis das IQWiG zu einer horizontalen und vertikalen Abstufung hinsichtlich der in der Matrix der Tabelle NT6 festgelegten relativen Risiken und der daraus für die Erreichung der jeweiligen Zielgröße zu unterschreitenden Schwellenwerte gekommen ist. Im Ergebnis führt das beschriebene Vorgehen dazu, dass eine Hierarchisierung der Zielgrößenkategorien hinsichtlich ihrer Bedeutung erfolgt, da sich beispielsweise in der Nutzenkategorie „erheblich“ die relativen Risiken für die Zielgrößen „Lebensqualität“ und „Überlebenszeit“ erheblich unterscheiden. Eine hinsichtlich ihres Effekts vergleichbare Verbesserung der Lebensqualität ist damit vereinfacht ausgedrückt weniger „wert“ als eine Verlängerung der Überlebensdauer. Eine solche Hierarchisierung hat der Verordnungsgeber bewusst nicht vorgenommen, er hat beispielsweise „eine moderate Verlängerung der Lebensdauer“ und eine „bedeutsame Vermeidung anderer Nebenwirkungen“ gleichwertig nebeneinandergestellt.

Die Festlegung eines Algorithmus beinhaltet eine Vielzahl ethischer Wertentscheidungen. Dem IQWiG fehlt wie dem G-BA eine klare demokratische Legitimation. Deshalb gilt es, durch wissenschaftliche Studien die Präferenzen von Patienten zu ermitteln. Durch geeignete Befragungstechniken und statistische Analyseverfahren sollte der Algorithmus so auf Patienten-Präferenzen basiert sein. Nur ein solches Vorgehen entspricht letztlich den internationalen Standards der evidenzbasierten Medizin und der Gesundheitsökonomie.

[4] Siehe z.B. Lerman J. Study design in clinical research: sample size estimation and power analysis. Can J Anaesth 1996(43): 184-191

f. Schwellenwerte müssen für abweichende Effektmaße angepasst werden.

In Abschnitt 3.3.3 B) (S. 18) des vorliegenden Dokuments wird festgelegt, dass die Konfidenzgrenzen für Risk Ratios auch auf Zielgrößen „Zeit bis Ereignis“ angewendet werden. Dies ist im Ticagrelor-Bericht bereits der Fall (Tabelle 13). Bei Zielgrößen „Zeit bis Ereignis“ handelt es sich üblicherweise um Hazard Ratios als Ergebnis von Survival-Analysen (Kaplan-Meier, Cox PH). Die Gleichsetzung der Konfidenzgrenzen impliziert die gleiche Verteilung der Schätzer für die Hazard Ratio und Risk Ratio auch bei kleinen Stichproben. Es fehlt eine Begründung für die Übertragbarkeit der Einstufungen von (nicht zeitpunktbezogenen) Risk Ratios auf (zeitpunktbezogene) Hazard Ratios.

Das Institut wird aufgefordert, für „Zeit-bis-Ereignis“ Variable eine eigenständige Bewertungsmethodik vorzulegen. Gegebenenfalls ist die Übertragbarkeit der Grenzwerte für Risk Ratios genauer zu begründen.

g. Die Skalentransformation für stetige Variablen im Rahmen von Responderanalysen führt zu einer beträchtlichen Power-Absenkung bei der Beurteilung von Effekten im Bereich der Lebensqualität.

Zur Messung der Lebensqualität ist der Einsatz von validierten Fragebögen (z.B. SF-36, SF-30) üblich. Die Einzelfragen werden zu Fragedimensionen (z.B. „physical functioning“, „role functioning“) zusammengefasst, die im Sinne des vorliegenden Dokuments „stetige oder quasi-stetige Zielgrößen“ sind (Abschnitt 3.3.3 C, S.19). Die Umwandlung in einer Responderanalyse mit vorgegebenem Cut-Off-Wert ist in diesem Fall wirklichkeitsfremd, da Cut-Off-Werte für die Fragedimensionen nicht existieren. Darüber hinaus führt der Informationsverlust, der mit der Transformation auf ein niedrigeres Skalenniveau verbunden ist, zu einem deutlichen Power-Verlust und zum Verlust der Kausalität zwischen Behandlung und Response[5] .

Das Institut wird aufgefordert, für stetige Variable eine Bewertungsmethodik vorzulegen, bei der der Informationsverlust minimiert wird. Dies betrifft insbesondere Lebensqualitätsstudien.

h. Differenzierung nach Erkrankung ermöglichen!

Die IQWiG-Methodik zur Operationalisierung des Ausmaßes des Zusatznutzens ist zu undifferenziert. Über alle Indikationen einheitlich festgesetzte Schwellenwerte können der Komplexität unterschiedlicher Erkrankungen nicht gerecht werden. So muss beispielsweise unterschieden werden zwischen chronisch belastenden, aber im Regelfall nicht tödlichen Erkrankungen und akuten, tödlich verlaufenden Krankheitsgeschehen. Der Forderung, den Schweregrad der Erkrankung zu berücksichtigen, trägt der Verordnungsgeber in § 5 Abs. 7

[5] Vgl. hierzu auch: Kleist, Peter. Wie sinnvoll sind Responderanalysen in klinischen Studien. Schweizer Med. Forum 2010: 169-171 und Altman, DG, Royston P. The cost of dichotomising continuous variables. BMJ 2006; 332: 1080

der AM-NutzenV bereits Rechnung. Hier heißt es wörtlich: „*Für Arzneimittel nach Absatz 3 sind das Ausmaß des Zusatznutzens und die therapeutische Bedeutung des Zusatznutzens* ***unter Berücksichtigung des Schweregrades der Erkrankung*** *gegenüber dem Nutzen der zweckmäßigen Vergleichstherapie wie folgt zu quantifizieren…*“

Beispiel: Feste Intervalle, wie in Tabelle NT1 festgelegt, führen dazu, dass Studiensettings mit einem großen Vorwissen benachteiligt werden. Ist der Behandlungseffekt bereits gut eingrenzbar, kann eine Studie genau geplant werden und wird zu einem signifikanten Ergebnis kommen, bei dem das Konfidenzintervall vermutlich knapp am Nulleffekt liegt. Dies kann selbst in Szenarien mit sehr gutem Behandlungseffekt geschehen - die Studien sind dann sehr klein. Es wäre dann unethisch, die Studie größer als notwendig zu machen. Allerdings würde die vorgeschlagene IQWIG Methode sich dann nachteilig auswirken, da - gemessen am Konfidenzintervall - nur ein geringer Zusatznutzen gezeigt werden kann. Dies widerspricht dem gesellschaftlichen Anspruch und den Forderungen der Ethik-Kommissionen, dass Studien unethisch sind, wenn mehr Patienten als notwendig eingeschlossen werden.

Auch in Bezug auf die Feststellung des Ausmaßes des Zusatznutzen kann ein Fehler erster (α-Fehler) und zweiter Art (β-Fehler) begangen werden. Dies soll an einem Beispiel verdeutlicht werden:

a) Einem Arzneimittel mit geringem Zusatznutzen wird fälschlicherweise ein beträchtlicher Zusatznutzen zugesprochen (α-Fehler).
b) Einem Arzneimittel wird ein nur geringer Zusatznutzen attestiert, obwohl tatsächlich ein beträchtlicher Zusatznutzen vorhanden ist (β-Fehler).

Wie bereits oben für die Feststellung des Zusatznutzen an sich vorgeschlagen, bedarf deshalb auch die Festlegung des Ausmaßes des Zusatznutzens einer indikationsspezifischen Differenzierung mit dem Ziel die negativen Folgen eines α-Fehler aber auch eines β-Fehler simultan zu minimieren.

Eine Weiterentwicklung der Methodik zur Bewertung des Zusatznutzens sollte aus Sicht des Verbandes:

- indikations- bzw. interventionsspezifisch sein,
- nicht primär aus statistischen Größen abgeleitet werden,
- auf messbaren Therapieerfolgen basieren,
- in den interdisziplinären Fachgremien intensiv diskutiert werden
- und Wertentscheidungen sollten durch Direkt-Betroffene und legitimierte Entscheidungsträger getroffen werden.

i. Berücksichtigung der Grenzen des Erkenntnisgewinns und Forderung nach einem rationalen Umgang mit Unsicherheit

Gerade bei schweren Erkrankungen ist im Rahmen der klinischen Prüfung beispielsweise oft ein Cross-Over-Design vorgeschrieben. Dies bedeutet, dass Patienten in der Vergleichsgruppe (also die Patienten im sogenannten Kontrollarm der Studie) das zu testende Arzneimittel erhalten müssen, falls sich Hinweise auf eine positive Wirkung des in der Studie zu testenden neuen Therapieansatzes ergeben. Cross-Over-Designs führen jedoch in der Regel zu statistisch schlechteren Ergebnissen. Dies führt dazu, dass unter Verwendung der IQWiG-Methodik zur Operationalisierung des Zusatznutzens bestimmte Arzneimittel systematisch niedriger bewertet werden.

Hinsichtlich der Beurteilung der Ergebnissicherheit von Bewertungen (Beleg, Hinweis, Anhaltspunkt) wird vom IQWiG ein methodischer Ansatz verfolgt, nach dem die gleichen Anforderungen für jedes zu bewertende Arzneimittel weitgehend unterschiedslos gelten sollen. Aus Sicht des Verbandes ist die IQWiG-Methodik dahingehend zu ergänzen, dass hinsichtlich der Anforderungen an die im konkreten Fall geforderte Ergebnissicherheit Adaptionen vorgenommen werden können, die sich bspw. an der Schwere der Erkrankung, den zur Verfügung stehenden Therapiealternativen sowie weiteren Parametern orientieren könnten, die fachlich und gesellschaftlich breiter zu diskutieren wären.

j. Lebensqualität und Patientenpräferenzen

Im Anhang „Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens“ werden die Kriterien gemäß AM-NutzenV um den Aspekt der gesundheitsbezogenen Lebensqualität ergänzt. Diese Ergänzung ist ausdrücklich zu begrüßen, da somit die Bedürfnisse der betroffenen Patienten ausreichend bei der Entscheidungsfindung berücksichtigt werden. In diesem Zusammenhang müssen dann konsequenterweise auch die im Dossier gelieferten Ergebnisse der Methoden CA und AHP (siehe 3.1.5) in der praktischen Bewertung des Zusatznutzens einfließen. Es ist daher auch im Anhang ein expliziter Verweis auf die multikriteriellen Methoden (AHP und DCE) aus Kapitel 3.1.5. erforderlich.

k. IQWiG-Methodik ist der falsche Ort für Werturteile

Das IQWiG gesteht folgendes selbst zu: *„Zu den Fragen, welche Effektstärken für die einzelnen Zielgrößen zu welcher Ausmaßkategorie führen und welche Effektmaße für diese Bewertung zu wählen sind, finden sich in der AM-NutzenV keine Angaben. Diese Fragen können prinzipiell nur bedingt methodisch beantwortet werden. Dennoch besteht die Notwendigkeit, das in den Dossiers dargelegte Ausmaß des Zusatznutzens zu bewerten (§ 7 Abs. 2, AM-NutzenV) und selbst Aussagen zum Ausmaß zu machen. Um hierbei zunächst die im weiteren Abwägungsprozess notwendigerweise zu treffenden Werturteile möglichst gering zu halten und diese explizit zu machen, bedarf es einer*

- *expliziten Operationalisierung, um ein transparentes und nachvollziehbares Verfahren sicherzustellen,*

sowie einer

- *abstrakten Operationalisierung, um größtmögliche Konsistenz zwischen den Nutzenbewertungen*

zu erzielen.“

Das IQWiG verlässt hier demnach nach eigenen Aussagen den Bereich der Wissenschaft und setzt selbst Normen, weil die gesetzlichen und untergesetzlichen Normen nicht ausreichend bestimmt sind. Das wissenschaftliche Institut IQWiG verfügt über keine demokratische Legitimation, Werturteile zu treffen und Normen zu setzen. An diesem grundsätzlichen Legitimationsdefizit ändert sich auch nichts unabhängig davon, wie transparent, nachvollziehbar und konsistent das IQWiG vorgeht. Zum Beispiel versteht das IQWiG unter „Konsistenz“ offenbar ein formalistisches Vorgehen nach dem Prinzip „one size fits all“. Aus Sicht des Verbandes ist jedoch gerade dies inkonsistent, weil Ungleiches (unterschiedlichste Erkrankungen mit ihren spezifischen Besonderheiten) immer gleich behandelt wird (fixe Schwellenwerte unabhängig von Besonderheiten einzelner Erkrankungen).

Aus Sicht des Verbandes ist deshalb zu fordern, dass die statistischen Schwellenwerte indikationsspezifisch bzw. interventionsspezifisch festgelegt werden. Auf die Ausführungen im Punkt h. wird verwiesen.

Zu 7.3.8 Meta-Analysen

Die Vorgehensweise zur Feststellung von Heterogenität der betrachteten Studien und die Vorgehensweise bei Vorliegen von Heterogenität sind extrem restriktiv.

Die in Kapitel 7.3.8 B) dargestellte Vorgehensweise zur Feststellung von Heterogenität von Studien innerhalb einer Metaanalyse ist so konzipiert, dass auch geringe Anzeichen für das Vorliegen von Heterogenität dazu führen, dass die Meta-Analyse insgesamt unterbleibt. Zur Feststellung wird ein Grenzwert von 0,2 beim Test von Cochran’s Q angesetzt. Liegt der p-Wert des Tests zwischen 0 und 0,2, unterbleibt bei nicht gleichgerichteten Effekten die Metaanalyse insgesamt. Das Institut wählt die Obergrenze des Bereichs von 0,1 bis 0,2, der im vorliegenden Dokument als „üblicherweise empfohlen“ zitiert wird. Andere Quellen (wie z.B. das Cochrane-Handbuch)[6,7] nennen einen p-Wert von 0,1, bis zu dem Heterogenität einer Studie festgestellt wird. Die Konsequenz der restriktiven Setzung des p-Werts besteht darin, dass auch Studienzusammensetzungen, die nach Beurteilung des I^2-Werts in die

[6] Cochrane handbook for systemativ reviews of interventions. Verion 5.1.0 (updated March 2011). [Zugriff 21.05.2013]. URL http://handbook.cochrane.org. Abschnitt 9.5.2 (Assessing and measuring heterogeneity)

[7] Higgins JPT, Thompson SG, Deeks JJ, Altman, DG. Measuring inconsistency in meta-analyses.BMJ 327(6):557-560

BPI Bundesverband der Pharmazeutischen Industrie e.V.

Friedrichstraße 148
10117 Berlin
www.bpi.de

Stellungnahme

Kategorie „Heterogeneity might not be important“ (I^2<=40%)[8] fallen, als „heterogen“ angesehen werden. In der Tat ist das schon ab einem I^2-Wert etwa 20-25% der Fall, bei hoher Anzahl von Studien auch schon für I^2-Werte unter 20%.

Das Institut wird aufgefordert, seine Methodik für schwach bzw. mäßig heterogene Metaanalysen zu präzisieren.

Berlin, 22.05.2013 MW/KG/VA

[8] „Rough guide for interpretation“ im erwähnten Abschnitt des Cochrane Handbook for Systematic Reviews of Interventions. [gleiche URL, gleiches Zugriffsdatum]

A.1.8 – Deutsche Diabetes Gesellschaft (DDG), Deutsche Gesellschaft für Innere Medizin (DGIM), Deutsche Gesellschaft für Kardiologie (DGK), Deutsche Krebsgesellschaft (DKG), Deutsche Gesellschaft für Verdauungs- und Stoffwechselkrankheiten (DGVS)

Berlin, 22. Mai 2013

Gemeinsame Stellungnahme

Deutsche Diabetes Gesellschaft (DDG)
Deutsche Gesellschaft für Innere Medizin (DGIM)
Deutsche Gesellschaft für Kardiologie (DGK)
Deutsche Krebsgesellschaft (DKG)
Deutsche Gesellschaft für Verdauungs- und Stoffwechselkrankheiten (DGVS)

zum Entwurf

„Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“ des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) vom 18.04.2013

Die fünf wissenschaftlichen Fachgesellschaften sehen den Stellenwert des Arzneimittelmarktneuordnungsgesetzes (AMNOG) sowie mit dem Gemeinsamen Bundesausschuss (G-BA) und dem Institut für Qualiät und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) die Bedeutung der entsprechenden Institutionen. Wir betrachten aber mit großer Sorge die Entwicklungen, die bei der Nutzenbewertung bestehender und neuer Medikamente erkennbar werden. Durch diese Entwicklungen wird die Versorgung der Patienten entscheidend beeinflusst und die ärztliche Therapiefreiheit teilweise eingeschränkt. Wir nehmen deshalb das Stellungnahmeverfahren des IQWiG zur Überarbeitung der „Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden 4.1“ zum Anlass, grundlegend zum Prozess der Frühen Nutzenbewertung nach § 35a SGB V Stellung zu nehmen. Die Stellungnahme richtet sich deshalb nicht nur an das IQWiG, sondern gleichermaßen an den G-BA. Die fünf Fachgesellschaften vertreten den allgemeinen internationalen Stand des medizinischen Wissens für die Therapie von Volkskrankheiten und Krankheitsbildern, die mit weitem Abstand die häufigsten medizinischen Todesursachen in Deutschland sind. Alle fünf Fachgesellschaften - in der Regel in Abstimmung mit europäischen Partnergesellschaften - geben, basierend auf der aktuellen wissenschaftlichen Evidenz, Leitlinien heraus.

Bisher sind aus unserer Sicht bei der Nutzenbewertung durch das IQWiG häufiger Diskussionen, unterschiedliche Ansichten, Streitigkeiten bis hin zu Fehleinschätzungen entstanden. Wir sehen durch Beachtung der folgenden fünf Punkte ein Potential zur qualitativen Verbesserung des Evaluationsprozesses:

1. **Festlegung der zweckmäßigen Vergleichstherapie**
2. **Formulierung der Fragestellung**
3. **Festlegung und Bewertung des Zusatznutzens**
4. **Einbindung von externen Gutachtern**
5. **Begründung bei Abweichung von Leitlinienempfehlungen**

Daher bitten wir bei der Überarbeitung des Methodenpapiers und der künftigen Arbeit des IQWiG um Berücksichtigung dieser fünf Punkte. Im Folgenden werden diese Punkte präzisiert:

Ad 1 Festlegung der zweckmäßigen Vergleichstherapie

Die Festlegung der Vergleichssubstanz muss durch den G-BA unter Einbeziehung der entsprechenden wissenschaftlichen Fachgesellschaften und unter Beachtung gültiger Behandlungsleitlinien erfolgen. Nur die wissenschaftlichen Fachgesellschaften können, wie vom Gesetzgeber vorgeschrieben, den aktuellen Stand des medizinischen Wissens vertreten und die spezifischen Aspekte, die bei der Behandlung von Krankheiten berücksichtigt werden müssen, einbringen. Zudem können sie die Sinnhaftigkeit des Vergleichs von medizinischen Therapiestrategien medizinisch-wissenschaftlich begründen und bewerten.

Ad 2 Formulierung der Fragestellung

Die Fragestellung bestimmt das Ergebnis. Wenn die Fragestellung nicht adäquat gewählt wird, kann jede methodisch noch so sorgfältige Analyse zu nicht plausiblen und für die Therapie unserer Patienten potentiell gefährdenden Einschätzungen führen. Auch hier empfehlen wir , dass die Expertise der entsprechenden wissenschaftlichen Fachgesellschaften zu Beginn eingeholt wird.

Ad 3 Festlegung und Bewertung des Zusatznutzens

Sind die zweckmäßige Verlgeichstherapie und die Fragestellung definiert, muss klinisch orientiert ein Zusatznutzen inhaltlich definiert und dann die entsprechenden Zielkriterien festgelegt werden. Auch dies ist ohne direkte aktive Einbindung der Fachgesellschaften nicht zielführend umzusetzen. Hier müssen ggf. allgemeine Regeln krankheits- und themenbezogen spezifiziert werden; dies ist ohne Fachgesellschaften ebenfalls nicht möglich.

Ad 4 Einbindung von externen Gutachtern

Im Gegensatz zu guter wissenschaftlicher Praxis wählt das IQWiG selbst die Gutachter aus und legt nicht dar, inwieweit gutachterliche Stellungnahmen Eingang in die Prozesse und ihre Ergebnisse finden. Die wissenschaftlichen Fachgesellschaften schlagen vor, dass aus ihnen heraus Gutachter gewählt bzw. dem IQWiG benannt werden, auf die dann transparent und effizient zugegriffen werden kann. Entsprechend der guten klinischen und wissenschaftlichen Praxis sowie auch international üblicher und bewährter Gepflogenheiten muss den Gutachtern begründet widergespiegelt werden, wie mit Ihrer Beurteilung seitens des IQWiG umgegangen wurde. Dies gilt auch für eingegangene Stellungnahmen. Zudem weisen die Fachgesellschaften darauf hin, dass es bei Krankheiten häufig mehrere Patientenorganisationen und Selbsthilfegruppen gibt. Das IQWiG hat den Begriff Patientenorganisation nicht definiert, auch hier können die Fachgesellschaften ihren Beitrag leisten.

Da die Frist von 4 Wochen für fundierte ehrenamtlich erstellte Stellungnahmen zu komplexen Themen in ungebührlicher Form kurz ist, bieten die vier Fachgesellschaften an, dass sie jeweils eine Kommission mit Mitgliedern gründen, die die Prozesse, Gespräche, Fragen, Klärungen und Diskussionen mit dem IQWiG frühzeitig und konstruktiv gestalten. Dies würde viele Probleme, Zeit, Diskussionen und aus unserer Sicht beklagenswerte und zum Teil nicht akzeptable Fehlentwicklungen vermeiden.

Ad 5 Begründung bei Abweichung von Leitlinienempfehlungen
Die evidenzbasierten Leitlinien der fünf Fachgesellschaften stellen ein wichtiges und in vielen Fällen auch wissenschaftlich evaluiertes Instrument zur Festlegung klinischer Standards dar. Die universitäre Lehre und die inhaltliche Ausrichtung der Facharztausbildung nehmen die Leitlinien der Fachgesellschaften zur Grundlage. Schließlich sind die Leitlinien Grundlage bei der Evaluation der Prozessqualität in der klinischen und ambulanten Versorgung. Wenn das IQWiG Empfehlungen ausspricht, die den Empfehlungen der Leitlinien widersprechen, sollte dies für die praktizierenden Ärzte nachvollziehbar sein und folglich der Widerspruch zu den Leitlinien wissenschaftlich begründet werden.

Wir glauben, dass nicht nur aus wissenschaftlichen und klinischen Gründen sowie Gründen der demokratischen Legitimierung die wissenschaftlichen Fachgesellschaften durch ein transparentes Verfahren bei den o.a. Punkten berücksichtigt werden müssen, sondern wir sind auch davon überzeugt, dass dies die politische Akzeptanz der Ergebnisse auch bei den Gesetzgebern, den Kostenträgern, den Leistungserbringern und insbesondere auch bei den betroffenen Patienten, ihren Angehörigen und bei der Bevölkerung erhöht.

Für die Deutsche Diabetes Gesellschaft (DDG)
PD Dr. Erhard Siegel (Präsident)
Prof. Dr. Dirk Müller-Wieland (Vors. der Kommission wissenschaftliche Stellungnahmen)

Für die Deutsche Gesellschaft für Innere Medizin (DGIM)
Prof. Dr. Dr. h.c. Ulrich R. Fölsch (Generalsekretär).

Für die Deutsche Gesellschaft für Kardiologie (DGK)
Prof. Dr. Christian Hamm (Präsident)
Prof. Dr. Heribert Schunkert (Vors. der Klinischen Kommission)

Für die Deutsche Krebsgesellschaft (DKG)
Dr. Johannes Bruns (Generalsekretär)

Für die Deutsche Gesellschaft für Verdauungs-und Stoffwechselkrankheiten (DGVS)
Prof. Dr. Markus M. Lerch (Präsident)
Prof. Dr. Stefan Zeuzem (Vorstand Leitlinien)

A.1.9 – Deutsche Gesellschaft für Gesundheitsökonomie e. V. (dggö)

dggö e.V. ● Universität Duisburg-Essen ● 45113 Essen

Geschäftsstelle

Universität Duisburg-Essen
Lehrstuhl für Gesundheitsökonomie
Schützenbahn 70, SM 108
D-45113 Essen

Telefon: 0201 183-3679
Fax: 0201 183-3716

E-Mail: geschaeftsstelle@dggoe.de
Internet: www.dggoe.de

Datum 19. Februar 2013

STELLUNGNAHME DER DGGO ZUM ENTWURF DES IQWIG „AKTUALISIERUNG EINIGER ABSCHNITTE DER ALLGEMEINEN METHODEN 4.0…“

Hiermit nimmt die Deutsche Gesellschaft für Gesundheitsökonomie Stellung zu Abschn. 3.1.5 („Zusammenfassende Bewertung“) und zum beabsichtigten neuen Abschnitt 3.3.3 „Nutzenbewertung von Arzneimitteln nach §35a SGB V“ des Allgemeinen Methodenpapieres und zu dem dazu gehörigen Anhang („Rationale…“):

Die dggö begrüßt es, dass das IQWiG die Möglichkeiten zu einer Nutzen- und Schadenskomponenten verrechnenden Bewertung auf der Grundlage einer Erfassung von individuellen Präferenzen reflektiert. Die dggö teilt allerdings nicht die Auffassung des IQWiG, dass in Deutschland anstelle der international weit verbreiteten Verwendung des QALY aufgrund ethischer und methodischer Probleme andere Bewertungsrahmen, wie die Effizienzgrenze, zu bevorzugen seien. Die intensive Erforschung und Diskussion des QALY-Konzepts hat die ihm zugrundeliegenden Werturteile, Annahmen und methodischen Defizite deutlich gemacht. Sie sind dadurch allgemein bekannt und können bei konkreten Bewertungen angemessen berücksichtigt werden. Die vom IQWiG vorgeschlagene alternative Me-

thode basiert ebenfalls auf (zum Teil gleichen) Werturteilen, Annahmen und methodischen Defiziten, die aber vom IQWiG weder transparent gemacht werden noch allgemein bekannt sind.

Die dggö begrüßt es gleichwohl, dass das IQWiG Verfahren der Präferenzmessung im Kontext der multidimensionalen Entscheidungsfindung, wie Analytic Hierarchy Process oder Conjonit-Analyse, in Betracht zieht. Beide Verfahren weisen allerdings erhebliche Unterschiede in der Zielsetzung, der Erhebung und Auswertung der Daten und besonders in der Interpretation derselben auf. Im Gegensatz zur Conjoint Analyse (Discrete-Choice Analyse), bei der die Befragten sich zwischen Alternativen entscheiden müssen und somit ihre Abwägungen aufdecken, ist der Analytic Hierarchy Process nicht mikroökonomisch fundiert. Eine Interpretation der Ergebnisse eines Analytic Hierarchy Process als Patientenpräferenzen im Kontext der Analyse des Patientennutzens ist somit fragwürdig. Es ist zu beachten, dass in der Fachdiskussion durchaus kritische Aspekte zur Anwendung dieser Verfahren artikuliert werden, die etwa Metrik, Anzahl und Art der einzubeziehenden Endpunkte, einzubeziehende Perspektiven und Charakteristika der zu befragenden Patientengruppen betreffen. Weiterhin ist zu beachten, dass die in diesem Kontext zu treffenden Werturteile (etwa der Frage, ob Versicherte oder Patienten befragt werden sollen) nicht vom IQWiG als wissenschaftlicher Einrichtung gefällt werden dürfen, sondern allenfalls vom G-BA, basierend auf einer öffentlichen Debatte, zu treffen wären.

Wie schon im Anhang zur Nutzenbewertung von Ticagrelor erstmalig durchgeführt, unternimmt das IQWiG in Abschn. 3.3.3 (und dem dazu gehörigen Anhang) den Versuch, einen Algorithmus zu entwickeln, mit dem für unterschiedliche Kategorien von Zielgrößen (z.B. Gesamtmortalität, schwerwiegende Symptome, Lebensqualität, vermiedene Nebenwirkungen) eine Zuordnung des Ausmaßes des Effektes zu den Ausmaßkategorien „erheblich“, „beträchtlich“ bzw. „gering“ der Arzneimittel-Nutzenbewertungsverordnung vorgenommen werden kann.

Die dggö stimmt mit dem IQWiG darin überein, dass es im durch Gesetz und Rechtsverordnung vorgegebenen Rahmen der Nutzenbewertung sinnvoll ist, einen Algorithmus zur Quantifizierung des Zusatznutzens zu entwickeln.

Die dggö sieht allerdings ein grundsätzliches Problem darin, dass hier im Gewand einer formalen „wissenschaftlichen“ Bewertung starke normative Aspekte eingehen, die im Grunde genommen des allgemeinen gesellschaftlichen Diskurs und der konkreten Erfassung der Präferenzen bedürften. Das IQWiG ist sich dieses schwerwiegenden Problems möglicherweise auch bewusst, wenn es ausführt (S. 31), es gelte, die „notwendigerweise zu treffenden Werturteile möglichst gering zu halten und diese explizit zu machen.“ Jenseits dieses Statements werden sodann allerdings ohne jedes Fundament in gesellschaftlichen Präferenzen Schwellenwerte „festgelegt“, bei denen für Kategorien von Outcomes Zusatznutzenkategorien relevant sein sollen. Zumindest halten wir es für erforderlich, die gewählten Schwellenwerte stärker inhaltlich zu begründen und ihre Einschränkungen, Probleme und Implikationen im Methodenpapier zu diskutieren.

Im Methodenpapier sollte sich das IQWiG darüber hinaus auferlegen, in den vorgelegten Nutzenbewertungen generell studienkompatible Angaben der verwendeten Schwellenwerte und Interpretation im Vergleich zum tatsächlichen Studienresultat vorzulegen. Sofern mehrere Nutzendimensionen in die Bewertung eingehen, sollte eine vergleichende Interpretation deren Schwellenwerte in studienkompatiblen Angaben erfolgen. Dies würde die Zielkonflikte transparent machen und es den Entscheidern, GBA und GKV-Spitzenverband, erleichtern, von den Empfehlungen abzuweichen und eigene Interpretationen vorzunehmen.

Jenseits dieser grundsätzlichen Kritik erachtet die dggö das vorgeschlagene Verfahren auch insoweit für problematisch, wie durch die Festschreibung der für eine Ausmaßkategorie des Zusatznutzens (z.B. "beträchtlich") geforderten Effekte über verschiedene Nutzendimensionen hinweg ein festes (Ausmaß-)Verhältnis zwischen diesen Nutzendimensionen festgelegt wird (z.B. 0,95 und 0,90 bei Mortalität und Lebensqualität), ohne dass dies inhaltlich begründet oder diskutiert wird.
Als problematisch sehen wir auch an, dass für völlig unterschiedliche Nutzendimensionen (wie Anteil der Patienten mit Nebenwirkung einer schweren Blutung und durchschnittliche Lebensqualität der Patienten) der gleiche Schwellenwert als Effektnachweis festgelegt

wird, ohne dass dies adäquat begründet oder gar empirisch auf Konsistenz und Plausibilität geprüft wird.

Essen, den 20. Mai 2013

Prof. Dr. Jürgen Wasem, Vorsitzender

A.1.10 – Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e. V. (DGHO)

Prof. Dr. med. Mathias Freund
Geschäftsführender Vorsitzender
Priv.-Doz. Dr. med. Diana Lüftner
Vorsitzende
Prof. Dr. med. Martin Wilhelm
Mitglied im Vorstand - Sekretär

DGHO e. V. • Alexanderplatz 1 • 10178 Berlin

Institut für Qualität und Wirtschaftlichkeit

im Gesundheitswesen (IQWiG)

Im Mediapark 8 (KölnTurm)

50670 Köln

Hauptstadtbüro der DGHO
Alexanderplatz 1 • 10178 Berlin
Tel. 030.27 87 60 89 - 0
Fax: 030.27 87 60 89 - 18
info@dgho.de

21. Mai 2013

Stellungnahme zur

Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1

Entwurf vom 18. April 2013

1. Zusammenfassung

Die Frühe Nutzenbewertung ist als ‚lernendes System' beschrieben worden. In diesem Sinn ist es sehr begrüßenswert, dass der jetzt vorgelegte Entwurf für die Allgemeinen Methoden 4.1 die bisher nur im Ticagrelor-Bericht dargestellte Operationalisierung der Feststellung des Ausmaßes des Zusatznutzens formal aufnimmt. Die Behebung weiterer Defizite wie ein Algorithmus zur Bewertung von Morbidität, für die Bewertung von Nebenwirkungen oder die Verankerung der Beteiligung Betroffener sollte zeitnah folgen.

2. Einleitung

Das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) wird vom Gemeinsamen Bundesausschuss (G-BA) mit der Erstellung von Berichten im Rahmen der Nutzenbewertung von Arzneimitteln beauftragt. Dies betraf bisher neu zugelassene Präparate, wird jetzt auf Arzneimittel des Bestandsmarktes ausgeweitet. Ausnahme sind

Arzneimittel für seltene Erkrankungen (Orphan Drugs), bei denen der Bericht vom G-BA erstellt wird. Das IQWiG wird bei Orphan Drugs mit einer Einschätzung der voraussichtlich zu behandelnden Patientenzahl und den voraussichtlich entstehenden Kosten beauftragt.

Die Nutzenbewertung von Arzneimitteln hat unmittelbaren Einfluss auf die Preise, mittelbar Einfluss auf die zukünftige Gestaltung klinischer Studien und potenziell auf die Versorgung der Bevölkerung mit wirksamen Arzneimitteln.

Die vom IQWiG angewandte Methodik ist im Schriftstück ‚Allgemeine Methoden' dargelegt, zuletzt gültig in der Version 4.0. Der jetzt vorgelegte Entwurf für die Version 4.1 ergänzt einige Punkte der bisherigen Version 4.0. In der Vergangenheit hatten wir in schriftlichen Stellungnahmen und im Rahmen von Anhörungen einige Defizite der IQWiG Berichte kritisiert, die auch Ursache abweichender Bewertungen von Arzneimitteln durch die Fachgesellschaft und durch den G-BA waren.

3. GRADE

Basis der Nutzenbewertung ist das Konzept der Evidenz-basierten Medizin. Diese ist auch Basis der Entwicklung von Leitlinien/Therapieempfehlungen und der Zulassung neuer Arzneimittel. Das umfangreichste und aktuelle System ist GRADE (Grading of Recommendations, Assessment, Development and Evaluation), seit 2004 entwickelt und inzwischen in mehr als 20 Publikationen differenziert dargestellt und begründet. Für wissenschaftliche medizinische Fachgesellschaften ist eine gemeinsame Grundlage zur Beurteilung der Qualität von Evidenz sinnvoll, nicht aber eine methodische Verselbstständigung der verschiedenen Anwendungsbereiche von Daten klinischer Studien. GRADE wird im aktualisierten Methodenpapier mehrfach zitiert. Für uns ist eine stärkere Orientierung der IQWiG Methodik an GRADE wünschenswert.

4. Operationalisierung der Feststellung des Ausmaßes des Zusatznutzens

Der Gesetzgeber hat die Bewertungskategorien der Nutzenbewertung (erheblich, beträchtlich, geringfügig, nicht quantifizierbar, kein Zusatznutzen, Nutzen geringer als Vergleichstherapie) vorgegeben. Aufgrund der Vielzahl von Einflussfaktoren in der Behandlung einer bestimmten Erkrankung, in einem bestimmten Stadium, in einer bestimmten Bevölkerungsgruppe und mit einem bestimmten Arzneimittel muss die Nutzenbewertung als ein Kontinuum angesehen werden. Die Kategorisierung vereinfacht zwar die Kommunikation von Ergebnissen und administrative Abläufe, schafft aber auch künstliche Trennschärfe.
Die Operationalisierung bei der Frühen Nutzenbewertung wurde für Ticagrelor erarbeitet. Zugrunde liegt eine Arbeit von Djulbegovic et al., 2008, mit einer umfangreichen Auswertung von Publikationen aus der Onkologie. Hier wurde ein „Relatives-Risiko-Quotient" (RR) von 0,5 bzw. 2,0 als Basis für die Definition eines erheblichen Zusatznutzens bei der Verlängerung der Gesamtüberlebenszeit festgesetzt. Nach dieser Festlegung würde in der Onkologie ein erheblicher Zusatznutzen bei 2% neuer Arzneimittel berechnet werden.

Die Methodik hat Schwächen. Sie liegen zum einen in der fehlenden Validierung, zum Beispiel anhand des langfristigen Einflusses eines Medikamentes auf die Gesundheit der behandelten Bevölkerungsgruppe. Der im Entwurf dieses aktualisierten Methodenpapiers

benutzte Begriff ‚verankert' suggeriert mehr Stabilität des Ansatzes als die Grundlagen hergeben.

Eine zweite Schwäche wird vom IQWiG selbst adressiert: „Die im Anhang A der Nutzenbewertung zu Ticagrelor [N7] aufgeführte Formel für den Zusammenhang des tatsächlichen Effekts und des Schwellenwerts ist unabhängig von den sonstigen Vorgaben und beruht auf dem Algorithmus, der in der Prozedur „Power" der Software SAS verwendet wird. In der entsprechenden Dokumentation für diesen Algorithmus [N9] wird auf die Arbeit von Fleiss et al. [N4] verwiesen. Ein Austausch mit Herrn Röhmel (damals Sprecher der Arbeitsgruppe Pharmazeutische Forschung der Deutschen Region der Internationalen Biometrischen Gesellschaft) sowie direkt mit dem Technical Support von SAS ergab, dass die Gültigkeit dieses Algorithmus offensichtlich nicht publiziert ist. Es stellte sich die Frage, welche tatsächlichen Effekte bei genauerer Berechnung notwendig sind, um mit einer hohen Wahrscheinlichkeit die jeweilige Ausmaßkategorie zu erreichen."

Die Übernahme der Operationalisierung aus dem Ticagrelor-Bericht in das Methodenpapier ist grundsätzlich zu begrüßen. Die Kritik am Algorithmus des Computerprogramms vertehen wir als eigenen Arbeitsauftrag.

5. Morbidität

Endpunkt in der Onkologie ist nicht nur Überleben, sondern auch Nicht-krank-sein. Wir haben wiederholt auf die Notwendigkeit einer differenzierten Betrachtung von Patienten-relevanten Endpunkten hingewiesen. In der jetzigen Aktualisierung wird das Thema nicht bearbeitet. Kriterien für die Bewertung von Morbidität müssen in der nahen Zukunft auf Ebene des G-BA und auf der gesundheitspolitischen Ebene diskutiert werden.

6. Nebenwirkungen

Nebenwirkungen neuer Arzneimittel sind ein entscheidendes Kriterium in der Risiko-Nutzen-Bewertung. Die allgemein verwendete CTCAE Klassifikation ist ein formales Gerüst zur Kategorisierung. Sie berücksichtigt nicht die mit der jeweiligen Nebenwirkung verbundene Belastung. Zum Beispiel ist die vom Patienten nicht gespürte und transiente Erhöhung eines Leberwertes im Grad 3 nicht mit einer Polyneuropathie im Grad 3 gleich zu setzen.

7. Sicherung der Beteiligung betroffener Patienten und behandelnder Ärzte

Ein Manko bisheriger IQWiG Berichte war das wiederholte Fehlen von betroffenen Patienten als Ratgeber. In Einzelfällen war auch kein ärztlicher Experte gefunden worden. Diese Defizite haben Einfluss auf die Übertragbarkeit der Berechnungen auf die reale Behandlungssituation und können eine Erklärung für die Diskrepanzen zwischen Ergebnissen der Berechnung und der endgültigen Bewertung sein.

Bei der Beteiligung namentlich genannter Fachgutachter ist zu überlegen, die von diesen Experten erbrachten Leistungen, i. e. schriftliche Beantwortung von Fragen, transparent darzustellen. Es bringt manchen Experten unter erheblichen Rechtfertigungsdruck, wenn sein Name auf einem Dokument erscheint, das in Details fachlich nicht korrekt ist und /oder dessen Schlussfolgerung in wesentlichen Punkten nicht mit seiner Einschätzung übereinstimmt.

8. Verständlichkeit

An Stellungnahmen der DGHO im Rahmen der Nutzenbewertung waren bisher mehr als 50 Kolleginnen und Kollegen beteiligt. Eine generelle Rückmeldung ist, dass die Methodik für den durchschnittlich versierten Arzt nicht transparent dargestellt wird.

Diese Stellungnahme wurde von Prof. Dr. Bernhard Wörmann in Kooperation mit Prof. Dr. Andreas Engert (Universitätsklinikum Köln, Klinik I für Innere Medizin, Köln) und Prof. Dr. Helmut Ostermann (Klinikum Großhadern der Ludwig-Maximilians-Universität, Medizinische Klinik III, München) erarbeitet.

Mit freundlichen Grüßen

Prof. Dr. med. Mathias Freund Geschäftsführender Vorsitzender	Priv.-Doz. Dr. med. Diana Lüftner Vorsitzende	Prof. Dr. med. Martin Wilhelm Mitglied im Vorstand - Sekretär

A.1.11 – Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS)

Aktualisierung der Allgemeinen Methoden Version 4.0

Gemeinsame Stellungnahme von GMDS und IBS-DR vom 15.05.2013

Autoren: Carsten Schwenke, Oliver Kuß, Uwe Siebert und Dieter Hauschke

Am 18.04.2013 wurde der Entwurf zur Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1 des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) zur Stellungnahme publiziert (IQWiG, 2013).

Die vorliegenden Entwürfe für Aktualisierungen und Ergänzungen des Methodenpapiers betreffen folgende Abschnitte:

- 2.1.1 Bericht
- 2.2.3 Review der Produkte des Instituts
- Neuer Abschnitt 3.1.4 Endpunktbezogene Bewertung
- 3.1.5 Zusammenfassende Bewertung (vorher Abschnitt 3.1.4)
- 3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V
- 7.3.8 Meta-Analysen
- Neuer Anhang: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens

2.1.1 Bericht

Kein Kommentar.

2.2.3 Review der Produkte des Instituts

Kein Kommentar.

Neuer Abschnitt 3.1.4 Endpunktbezogene Bewertung

Seite 8 letzter Absatz [1]:

"Ein wichtiges Kriterium zur Ableitung von Aussagen zur Beleglage ist die Ergebnissicherheit. [...] Falls die zugrunde liegenden Daten dies zulassen, lässt sich die statistische Unsicherheit als Standardfehler bzw. Konfidenzintervall von Parameterschätzungen quantifizieren und beurteilen (Präzision der Schätzung)."

Hier bietet es sich an, eingangs die beiden besprochenen Unsicherheitsarten (quantitative und qualitative Unsicherheit) auch mit den gängigen epidemiologischen Begriffen zu definieren.

Ergebnisunsicherheiten können durch zwei verschiedene Fehler auftreten:

1. Zufälliger Fehler verursacht durch eine Stichprobenziehung (statistische Unsicherheit)
2. Systematischer Fehler (Bias, systematische Verzerrung) im Wesentlichen beeinflusst durch Studiendesign bzw. korrekte Analyse. Hierzu gehören Informationsfehler, Selektionsfehler und Confounding.

GMDS Geschäftsstelle

Industriestraße 154
D-50996 Köln

Telefon: +49 (02236) 33 19 958
Telefax: +49 (02236) 33 19 959

E-Mail: info@gmds.de
Internet: www.gmds.de

IBS-DR Geschäftsstelle

Heike Krubert
c/o Institut für Biometrie, Epidemiologie und Informationsverarbeitung
Tierärztliche Hochschule
Hannover
Bünteweg 2
D-30559 Hannover

Telefon: +49 (0) 511 953 79 51
Telefax: +49 (0) 511 953 79 74

E-Mail: biometrische-gesellschaft@tiho-hannover.de

"Das Institut verwendet die folgenden drei Kategorien zur Graduierung des Ausmaßes der qualitativen Ergebnissicherheit auf Einzelstudien- und Endpunktebene:

- **hohe qualitative Ergebnissicherheit**: Ergebnis einer randomisierten Studie mit niedrigem Verzerrungspotenzial.
- **mäßige qualitative Ergebnissicherheit**: Ergebnis einer randomisierten Studie mit hohem Verzerrungspotenzial.
- **geringe qualitative Ergebnissicherheit**: Ergebnis einer nicht randomisiert vergleichenden Studie. " [1]

Daraus leitet sich ab, dass Ergebnisse aus nicht-interventionellen vergleichenden Studien wie prospektive Kohortenstudien oder auch Register, in denen neben dem zu bewertenden Arzneimittel auch Vergleichspräparate einbezogen werden, zur Nutzenbewertung herangezogen werden können. Es sind also nicht zwingend ausschließlich nur randomisierte kontrollierte Studien (RCT) zur Bewertung heranziehbar, sondern auch zusätzliche Evidenz aus anderen Studien, solange sie vergleichend ausgeführt sind.

Anforderung an eine Studie zum Nachweis eines Belegs:
"Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und ihre Ergebnisse besondere Anforderungen zu stellen" mit Referenz auf die Richtlinie "European Medicines Agency. Points to consider on application with: 1. meta-analyses; 2. one pivotal study [online]. 31.05.2001". Diese Richtlinie wurde erarbeitet für nichtonkologische Indikationen, in denen regelhaft mindestens zwei RCT gefordert werden, um eine Marktzulassung für ein neues Arzneimittel zu erhalten. In onkologischen Indikationen und bei Indikationserweiterungen ist dagegen häufig nur eine pivotale Studie gefordert.
Es stellt sich nun insbesondere für die frühe Nutzenbewertung gemäß § 35a SGB V die Frage, ob in den Indikationen, in denen die Zulassung auf einer Studie basiert und sogenannte Megatrials nicht möglich sind, höchstens ein "Hinweis" erreicht werden kann. Somit wäre die Ergebnissicherheit in diesen Indikationen beschränkt auf drei Kategorien (Hinweis, Anhaltspunkt, keines von beiden). Hier wären Überlegungen wünschenswert, wie für diese Indikationen ebenfalls vier Kategorien ermöglicht werden können, insbesondere in der frühen Nutzenbewertung, zu welcher nur die Zulassungsstudien vorliegen. Zudem wäre eine Klärung der Regeln wünschenswert, wann eine einzige Studie einen Beleg ermöglicht, da sich in den bisherigen Bewertungen nur ein fünfstelliger Stichprobenumfang als hinreichend dargestellt hat. In der Bewertung von Ticagrelor schreibt das IQWiG dazu "Wegen der besonderen Güte und der ausreichenden Größe der PLATO-Studie konnten aus den Daten Belege, z. B. für einen Zusatznutzen, abgeleitet werden", allerdings ohne Beschreibung, was unter "besonderer Güte" zu verstehen ist [2]. Hier wäre ein Hinweis auf die Richtlinie zu Meta-Analysen sinnvoll, in der die entsprechenden Qualitätsmerkmale benannt sind.

In der Tabelle 2 zu den Anforderungen an die Beleglage für die unterschiedlichen Aussagesicherheiten beim Vorliegen von Studien derselben qualitativen Ergebnissicherheit wird unter Hinweis und Anhaltspunkt aufgeführt:

Hinweis	≥2	hoch	heterogen mäßig gleichgerichtet
Anhaltspunkt	≥2	mäßig	heterogen mäßig gleichgerichtet

Da zwei Studien mit gleichgerichteten Effekten immer deutlich gleichgerichtet sind (siehe Entwurf, Seite 10), sollte es lauten:

Hinweis	≥**3**	hoch	heterogen mäßig gleichgerichtet
Anhaltspunkt	≥**3**	mäßig	heterogen mäßig gleichgerichtet

3.1.5 Zusammenfassende Bewertung (vorher Abschnitt 3.1.4)

"Eine Möglichkeit der gleichzeitigen Würdigung von Nutzen und Schaden ist die Gegenüberstellung der endpunktbezogenen Nutzen- und Schadenaspekte. Dabei werden die Effekte auf alle Endpunkte (qualitativ oder semiquantitativ) gegeneinander abgewogen mit dem Ziel, zu einer endpunktübergreifenden Aussage zum Nutzen bzw. Zusatznutzen einer Intervention zu kommen."

Es wäre wünschenswert, eine genauere Spezifizierung zu bekommen, nach welchen Kriterien diese Nutzen- zu Schadenabwägung durchgeführt wird. Sinnvoll wäre eine medizinische Abwägung auf Basis der Intensität/Schwere des Schadens gegenüber der Größe des Nutzens. Diese Abwägung sollte allerdings nur unter Miteinbeziehung der jeweiligen Indikation und deren Parameter wie z.B. Lebensbedrohung und Schwere der Krankheitssymptome erfolgen unter Einbeziehung indikationsspezifischer Experten erfolgen. Hier wäre eine genauere Beschreibung der Methodik und des Vorgehens wünschenswert.

Der folgende Abschnitt zu den Methoden der Präferenzenintegration ist neu:
S. 13 [1]: *„Häufig werden sogenannte Nutzwerte für Gesundheitszustände erhoben, die die von den Befragten positiv wie negativ empfundenen Aspekte in einer Indexzahl ausdrücken sollen. Unter Integration der Dauer der entsprechenden Gesundheitszustände können diese Nutzwerte bspw. in sogenannte qualitätsadjustierte Lebensjahre (QALYs = Quality-Adjusted Life Years) überführt werden. Aufgrund der ethischen und methodischen Probleme gerade der häufig verwendeten QALYs [122,137,138,520] sollten alternative Verfahren der multikriteriellen Entscheidungsfindung oder der Präferenzerhebung angewendet werden. Dazu zählen u. a. der Analytic Hierarchy Process (AHP) und die Conjoint-Analyse (CA).“*

Die Zusammenführung verschiedener Outcome-Dimensionen wie Lebensdauer, Morbidität und gesundheitsbezogener Lebensqualität ist ein komplexes Thema und es gibt weder ein einziges allgemeingültiges krankheitsübergreifendes Konzept, noch ist es einfach so, dass eines der genannten Instrumente durch „ethische und methodischen Probleme“ zu disqualifizieren wäre. Auch ist es nicht so, dass sich diese Instrumente gegenseitig ausschließen, sondern sie können komplementär eingesetzt werden. Die Empfehlungen hierzu sollten wissenschaftlich nüchtern gegeben werden. Dabei ist insbesondere zu berücksichtigen, dass der Entscheider auf der einen Seite unterschiedliche Dimensionen wie Lebensdauer und Lebensqualität bzw. unterschiedliche Dimensionen der Lebensqualität getrennt betrachten möchte und auf der anderen Seite diese verschiedenen Dimensionen integriert bzw. miteinander verrechnet werden müssen, wenn die Abwägung systematisch und nachvollziehbar stattfinden soll. Gerade die Getrennthaltung verschiedener Dimensionen wird dann schwierig nach-

zuvollziehen, wenn das IQWiG diese Dimensionen dann in einem nicht weiter definierten gewichteten Score wieder zusammenführen muss.
Siehe dazu im IQWiG Dokument auf S. 13 [1]: *„Eine weitere Möglichkeit der gleichzeitigen Würdigung besteht darin, die verschiedenen patientenrelevanten Endpunkte zu einem einzigen Maß zu aggregieren. In diesem Fall würden die Aussagen des Instituts für jeden einzelnen patientenrelevanten Endpunkt gewichtet z. B. in einen Summenscore einfließen."*
Bei einer solchen Zusammenführung müssen ähnliche Annahmen gemacht werden, wie sie dem QALY zugrunde liegen, z.B. dem Abwägen zwischen Lebensdauer und Lebensqualität (s. dazu auch [5-7]). Es wäre wünschenswert, wenn das IQWiG an dieser Stelle keine inhaltliche Vorentscheidungen für Methoden trifft, sondern explizit festhält, dass die Auswahl der Integrationsmethode von der Indikation und Fragestellung abhängt und hierfür grundsätzlich alle Verfahren (QALY, AHP und CA) in Frage kommen, und kontextspezifisch geprüft werden müssen.
Insgesamt erscheint die Argumentation zum QALY-Konzept gegenüber dem ansonsten sehr seriös und wissenschaftlich gehaltenen Aktualisierungsdokument vergleichsweise unwissenschaftlich, was dem Gesamtdokument schadet. So wird auf S. 13 a) eine wissenschaftlich fehlerhafte Argumentation vorgenommen und b) werden die Argumente mit Literaturstellen belegt, die die Argumente nicht durchgängig stützen. Dies ist besonders problematisch, da beim Leser der (falsche) Eindruck entsteht, dass die zitierten Literaturstellen sinngemäß den Text im Methodenhandbuch belegen.
So handelt es sich bei den vier zitierten Literaturstellen (122, 137, 138, 520) vorwiegend um den Einsatz von Präferenzmaßen im Kontext der Gesundheitsökonomie bzw. Ressourcenallokation, nicht um den Kontext der Nutzen-Schaden-Abwägung.
Ein Beispiel: Das Dokument des Deutschen Ethikrats. Nutzen und Kosten im Gesundheitswesen: zur normativen Funktion ihrer Bewertung; Stellungnahme. Berlin: Deutscher Ethikrat; 2011. URL: http://www.ethikrat.org/dateien/pdf/stellungnahme-nutzen-und-kosten-im-gesundheitswesen.pdf.
Der Kontext und Anlass der Stellungnahme des Deutschen Ethikrats ist eher die Nutzen-Kosten-Bewertung des IQWiG. Dort wo es im Dokument um die reine Nutzenbewertung geht (Kapitel 3.1.3 Das Nutzenmaß QALY als Produkt aus Lebensqualität und Lebensdauer), wird das QALY-Konzept neutral mit seinen Vorzügen und methodischen Annahmen beschrieben. Auf S. 38 wird explizit auf die Plausibilität der QALY-Methode eingegangen, am Beispiel der „oftmals nebenwirkungsreichen Maßnahmen in der Tumorbehandlung wie Chemo- oder Strahlentherapie. Hier ist eine –manchmal fragliche – Verlängerung der Lebenserwartung mit einer nicht selten deutlich geminderten subjektiven ‚Erlebensqualität' der verbliebenen Lebenszeit verbunden." (S. 37)
Die Literatur zur Nutzen-Schaden-Abwägung, die sich durchaus vorsichtig positiv für das QALY-Konzept ausspricht, insbesondere, wenn das QALY-Konzept innerhalb einer Indikation angewendet wird, fehlt im IQWiG-Dokument. Siehe beispielsweise:

- Puhan MA, Singh S, Weiss CO, Varadhan R, Boyd CM. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. BMC Med Res Methodol. 2012 Nov 19;12:173. doi: 10.1186/1471-2288-12-173.
- Garrison LP Jr, Towse A, Bresnahan BW: Assessing a structured, quantitative health outcomes approach to drug risk-benefit analysis. Health Aff (Millwood) 2007, 26(3):684-695.

Im IQWiG-Dokument (S. 13f) werden die Methoden Analytic Hierarchy Process und Conjoint-Analyse kurz beschrieben. Anwendungen der Methoden Analytic Hierarchy Process (AHP) und Conjoint-Analyse (CA) haben gezeigt, dass diese Konzepte nicht ohne Weiteres auf den Kontext der (Zusatz-) Nutzenbewertung anwendbar sind bzw. ebenfalls auf multiplen Annahmen beruhen, die in vielen

Entscheidungssituationen nicht erfüllt sind. Wünschenswert wäre deshalb eine kurze Beschreibung dazu, wie diese Instrumente im Kontext von Nutzenbewertungen durch das IQWiG eingesetzt werden sollen, und mit welchen Limitationen sie behaftet sind.

3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V

In der Operationalisierung der Feststellung des Ausmaßes des Zusatznutzens werden im 1. Schritt die Kriterien zur Ableitung von Aussagen zur Beleglage nach Abschnitt 3.1.4 angewendet. Hier stellt sich die Frage, ob die Regeln der Nutzenbewertung auch für die frühe Nutzenbewertung gemäß § 35a SGB V gelten können oder ob hier andere Regeln gelten sollten, da die Voraussetzungen von Nutzenbewertung und früher Nutzenbewertung durch z.B. die unterschiedlich lange Präsenz der Produkte auf dem Markt und dementsprechend unterschiedliche Verfügbarkeit verwendbarer Evidenz als sehr unterschiedlich einzustufen sind.

Im zweiten Schritt wird das Ausmaß des Nutzens festgestellt. Hierbei ist zu berücksichtigen, dass, wie unter 3.1.4 und für Schritt 1 diskutiert, in der frühen Nutzenbewertung sehr häufig nur wenige oder sogar nur eine verwendbare Studie vorliegt und damit das zweiseitige 95% Konfidenzintervall des Behandlungseffekts auf wenigen Studien basiert, wohingegen in der Nutzenbewertung häufig viele Studien zur Verfügung stehen, so dass der Behandlungseffekt präziser geschätzt werden kann. Daraus ergibt sich wie schon für die Ergebnissicherheit die Frage, ob in der frühen Nutzenbewertung dieselben Kriterien herangezogen werden sollten, um den Ergebnissen die Kategorien des Ausmaßes des Zusatznutzens zuzuordnen.

Schritt 3 ist dagegen äquivalent für die Nutzen- wie auch frühe Nutzenbewertung verwendbar, da es sich hierbei um die Nutzen- zu Schadenabwägung handelt, die in beiden Bewertungen gleich verwendet werden können.

C) Stetige oder quasi-stetige Zielgrößen mit jeweils vorliegenden Responderanalysen

Für stetige und quasi-stetige Zielgrößen werden Responderanalysen auf Basis validierter bzw. etablierter Responsekriterien verlangt. Damit wird verlangt, dass Zielgrößen mit einem hohen Informationsgehalt dichotomisiert werden sollen, wodurch ein Großteil der Information der Daten verloren geht. In diversen Indikationen wird von der Europäischen Zulassungsbehörde ein metrischer Endpunkt als primärer Endpunkt verlangt, auf dessen Basis die Stichprobenplanung der Studie basiert (z.B. Diabetes mellitus [3] mit dem "change from baseline in HbA1c"). Zudem werden Morbiditäts- und Lebensqualitätsdaten häufig in Form eines "visual analoque scale" dargestellt, für die dann der "change from baseline" bestimmt wird, z.B. für die Bewertung von Änderungen des Symptoms Schmerz (siehe auch [4]).

In der Tat lassen sich absolute Differenzen auf Basis metrischer Endpunkte nicht über Indikationen standardisieren. Möglich wäre dies zum Beispiel mit relativen Effektgrößen wie Cohen's d. Es bleibt aber die Frage, inwieweit sich diese dann über Indikationen hinweg vergleichen lassen in Bezug auf statistischer Signifikanz verknüpft mit klinischer Relevanz und welche Grenzen dann für die Bewertung des Ausmaßes des Zusatznutzens verwendet werden sollen. Eine Dichotomisierung und der damit verbundene Verlust vorhandener Information kann aber nicht die Lösung für diese Herausforderung sein. Sicher macht es Sinn, für metrische Daten sowohl den absoluten wie auch relativen Ef-

fekt zu betrachten, wenn denn letzterer möglich ist (d.h. eine adäquate Responderdefinition vorliegt, wie z.B. in der Indikation rheumatoide Arthritis. Zudem muss (indikationsspezifisch) definiert werden, wie dann das Ausmaß des Zusatznutzens bestimmt wird.

Abschnitt 7.3.8

Die Änderungen im Kapitel 7.3.8 Meta-Analysen betreffen im Wesentlichen nur einen neu eingefügten Abschnitt zur Angabe von Prädiktionsintervallen im Unterabschnitt B) Heterogenität, dieser entspricht aus unserer Sicht dem gegenwärtigen Forschungsstand. Leider scheint sich der Inhalt der gemeinsamen Stellungnahme von GMDS und IBS-DR vom Februar 2012 zum Kapitel 7.3.8 noch nicht in der Aktualisierung niedergeschlagen zu haben. Wir ermutigen das IQWIG besonders, die routinemäßige Anwendung der Standardmodelle mit zufälligen und festen Effekten zu überdenken, wo doch die Nachteile dieser Methoden altbekannt sind und zahlreiche Methoden zu deren Verbesserung vorgeschlagen worden sind.

Köln/Hannover, 15. Mai 2013

Mit freundlichen Grüßen

Prof. Dr. Heike Bickeböller
- Präsidentin der GMDS -

Dr. Jürgen Kübler
- Präsident der IBS-DR -

Anhang: Literatur

Literatur

[1] IQWiG (2013): *Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1 vom 18.04.2013.* Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Köln.

[2] IQWiG (2011): *Ticagrelor – Nutzenbewertung gemäß § 35a SGB V*. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Köln, IQWiG-Berichte – Jahr 2011 Nr. 96.

[3] CHMP (2012): *Guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus*. Committee for medicinal products for human use (CHMP), CHMP/EWP/1080/00 Rev. 1

[4] CHMP (2006): *Reflection paper on the regulatory guidance for the use of health related quality of life (HRQL) measures in the evaluation of medicinal products*. Committee for medicinal products for human use (CHMP), EMEA/CHMP/EWP/139391/2004

[5] Marckmann G, Siebert U. Prioritäten in der Gesundheitsversorgung: Was können wir aus dem „Oregon Health Plan“ lernen? Dtsch Med Wochenschr. 2002;127:1601-4

[6] Marckmann G, Siebert U. Kosteneffektivität als Allokationskriterium in der Gesundheitsversorgung. Zeitschrift für medizinische Ethik. 2002;48:171-190

[7] Marckmann G. Priorisierung im Gesundheitswesen: Was können wir aus den internationalen Erfahrungen lernen?, in: Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 2009;103(2):85-91.

A.1.12 – Deutsche Krankenhausgesellschaft e. V. (DKG)

Stellungnahme der Deutschen Krankenhausgesellschaft

zum Entwurf des IQWiG

„Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“

Initial möchten wir darauf hinweisen, dass wir die Kritikpunkte unserer ausführlichen Stellungnahme aus dem Jahr 2011 zu all denjenigen Punkten aufrechterhalten möchten, die bisher unberücksichtigt blieben. Um Redundanzen zu vermeiden, werden sie in dieser Stellungnahme nicht nochmals aufgeführt. Es handelt sich hier teilweise um Aspekte, auf die wir bereits seit mehreren Jahren aufmerksam machen. Beispielsweise würden wir weiterhin eine obligatorische mündliche Anhörung zu den Stellungnahmen begrüßen, da hierbei Unklarheiten und Rückfragen zu den Stellungnahmen besser als auf dem Schriftwege geklärt werden könnten. Die mündlichen Anhörungen möchten wir als konstruktiven Dialog verstehen.

Zu Kapitel 2.2.3 Review der Produkte des Instituts

„Darüber hinaus kann im Verlauf von der Erstellung von Berichten und z. T. auch von Gesundheitsinformationen ein externes Reviewverfahren als optionaler weiterer Schritt der Qualitätssicherung durchgeführt werden.“

Unabhängig von der fehlenden Begründung, warum bei der Verfassung von Vorberichten zukünftig ein externes Reviewverfahren nur noch optional und nicht mehr wie bisher regelmäßig stattfinden soll, wird aus unserer Sicht hierdurch eine wichtige Chance vertan, die methodische Kompetenz des Institutes durch breitere medizinisch-wissenschaftliche Kompetenz zu ergänzen und damit zu sachgerechten Bewertungen zu kommen. Sie adressieren zu Recht in dem letzten Abschnitt dieses Kapitels, dass das externe Qualitätssicherungsverfahren unter Beteiligung von Reviewern ein offenes und unabhängiges Reviewverfahren gewährleistet. Daher erschließt sich für uns umso mehr nicht, warum das externe Reviewverfahren nur noch als optional dargestellt wird.
Es werden weiterhin die Schritte der internen Qualitätssicherung nicht genauer benannt. Diesen Punkt hatten wir bereits in unserer letzten Stellungnahme angemerkt. In Ihrer Antwort hierzu wurde erklärt, dass „ein Institutsprodukt in der Folge zahlreichen Reviewverfahren unterworfen wird, u.a. auch einem externen Review, welches zusätzlicher Bestandteil der Qualitätssicherung ist.“ Unter dem Gesichtspunkt, dass nun ein wichtiger Schritt der Qualitätssicherung gestrichen werden soll, erscheint es umso wichtiger, die einzelnen Schritte des internen Qualitätssicherungsverfahrens transparent darzustellen.
Wir begrüßen es, dass für interne und externe Reviewer die Darlegung potenzieller Interessenkonflikte erforderlich ist. Diese Offenlegung fordern Sie auch von Stellungnehmenden, die im Verlauf eines Projekts eine institutionsgebundene Stellungnahme abgeben und ggf. an einer mündlichen Anhörung teilnehmen. Hierbei stellt sich für uns aber die Frage, warum die an einem Bericht beteiligten Mitarbeiter des IQWiG nicht verpflichtet werden, ihre Interessenkonflikte für die Öffentlichkeit in gleicher Weise im Bericht darzulegen.

„Die Auswahl der internen und externen Reviewer erfolgt primär auf Basis ihrer methodischen und/oder fachlichen Expertise.“

Wie auch bereits in der letzten Stellungnahme angemerkt, sollten die Auswahlkriterien für externe Reviewer und Sachverständige klarer dargestellt werden und es sollte insbesondere Wert auf fachlich-medizinische Expertise gelegt werden. Die Vergangenheit hat gezeigt, dass teilweise die Auswahl Ihrer externen Sachverständigen mangels adäquaten klinischen und wissenschaftlichen Hintergrunds nicht nachvollziehbar waren.

Darüber hinaus möchten wir anmerken, dass die von Ihnen genutzten Bezeichnungen *„interner Reviewer",* „*externer Sachverständiger*" und „*externer Reviewer*" definitorisch nicht klar voneinander abgegrenzt sind und die mangelnde Abgrenzung ihrer Aufgabenbereiche für Verwirrungen sorgt. Bei den Mitarbeitern und Mitarbeiterinnen des IQWIG handelt es sich vorrangig um Fachleute für evidenzbasierte Medizin und Health Technology Assessment, d.h. um methodische Expertise. Themenspezifische originäre Forschung mit entsprechenden Primärstudien werden vom IQWIG nicht erarbeitet, so dass es nicht nachvollziehbar ist, nach welchen Kriterien - außer methodischen - medizinisch-inhaltliche Wissenschaftsexpertise über das interne Review abgedeckt werden kann.

Zu Kapitel 3.1.4 Endpunktbezogene Bewertung

„…drei Kategorien zur Graduierung des Ausmaßes der qualitativen Ergebnissicherheit auf Einzelstudien- und Endpunktebene" (S.9) und Tabelle 2 (S. 11)

Als Cut-Off für die drei Kategorien zur qualitativen Ergebnissicherheit der Einzelstudien werden von Ihnen RCT gewählt und diese dann nochmal unterteilt in RCT mit niedrigem und hohen Verzerrungspotenzial. Angesichts der Themenvielfalt, die das IQWIG bearbeitet, ist dieses Vorgehen insofern sehr rigide und pauschal, da es auch klinische Fragestellungen gibt, zu denen nur sehr wenige oder gar keine RCT vorliegen, so dass auch weitere Evidenzklassen herangezogen werden sollten, um den allgemein anerkannten Erkenntnisstand abzubilden, der bei Ihrem Vorgehen immer in die niedrigste Kategorie fällt, gleichwohl dazu weitere Differenzierungen in der Verfahrensordnung des G-BA wie auch in internationalen Klassifikationsschemata für Evidenzlevel existieren (z.B. Oxford Centre for Evidence-based Medicine - Levels of Evidence). Diese Situation ist z. B. für kleine Subgruppen oder seltene Krankheitsentitäten oder in der Bewertung diagnostischer Verfahren zu antizipieren. Entsprechend kritisch erachten wir Tabelle 2, da sämtliche Evidenz unterhalb RCT – völlig unabhängig von der jeweiligen Fragestellung – ebenfalls in die niedrigste Kategorie fällt, also maximal einen Anhaltspunkt zur Nutzenbewertung liefern kann. Es fehlen weiterführende empirische Begründungen und Referenzierungen zur internationalen Literatur.

Kapitel 7.3.8 Meta-Analysen

„A) Allgemeines" (S. 20, 21)

In Bezug auf Meta-Analysen liegt laut aktuellem „Cochrane Handbook for Systematic Reviews Version 5.1.0" Random-Effects-Modellen die Annahme zugrunde, dass es sich um systematische Unterschiede bei den Einzeleffekten der eingehenden

Studien handelt (Heterogenität). Den Fixed-Effects-Modellen hingegen liegt die Annahme zugrunde, dass es keine systematischen Unterschiede gibt, d.h. die Unterschiede rein zufällig sind und von guter Vergleichbarkeit auszugehen ist (Homogenität). Die Unterscheidung zwischen zugrundeliegender Annahme und Bezeichnung der Modelle wird in Ihren Ausführungen nicht deutlich und Sie bedienen sich nicht der gebräuchlichen internationalen Terminologie, so dass es anhand Ihrer deutschen Übersetzungen zur sprachlichen Verwirrung und Unklarheiten kommen kann. Nach dem Cochrane Handbook hat es in Meta-Analysen zudem grundsätzlich Vorrang, mit Fixed-Effects-Modellen zu arbeiten, die von Homogenität der Ergebnisse der Einzelstudien ausgehen, um nicht „Äpfel mit Birnen" zu vergleichen. Zur Prüfung der Stabilität des Gesamteffekts können im Rahmen von Sensitivitätsprüfungen dann auch Random-Effects-Modelle untersucht werden. Darüber hinaus ist es in heterogenen Studiensituationen eine von verschiedenen möglichen Strategien, mit Random-Effects-Modellen zu arbeiten. Warum das IQWIG aber nun vorrangig mit Random-Effects-Modellen anstatt mit den zuverlässigeren Fixed-Effects-Modellen arbeiten wird, ist demzufolge nicht nachvollziehbar. Zudem sind die Gesamteffekte in Random-Effects-Modellen oft unpräziser und erreichen schwerer das Signifikanzniveau. Oder liegt hier insgesamt nur ein sprachliches Missverständnis vor?

„B) Heterogenität" (S. 21 ff.)

In Bezug auf Kapitel 7.3.8 haben wir bereits in der letzten Stellungnahme zu Ihrem Methodenpapier 4.0 angemerkt, dass nicht so sehr die statistische, sondern vielmehr die inhaltliche Untersuchung der Heterogenität im Vordergrund stehen sollte. Heterogenität, d.h. systematische Unterschiede in den Einzelergebnissen der Studien können in einer Vielzahl möglicher Unterschiede zwischen den in einem systematischen Review eingehenden Studien begründet sein (wie z. B. in unterschiedlichen Populationen, Interventionen, Settings etc.) Entsprechende Ausführungen sind auch im aktuellen Cochrane Handbook zu finden.

Ihre grundsätzliche Festlegung des Signifikanzniveaus bei $p \geq 0{,}1$ bis 0,2 für Homogenität in Heterogenitätstests ist vor dem Hintergrund, dass im Cochrane Handbook selbst ein $p \geq 0{,}1$ nur in Ausnahmefällen als sinnvoll angesehen wird, während der übliche Wert $p \geq 0{,}05$ beträgt, nicht nachvollziehbar. Mit diesem Vorgehen wird die Homogenitätsannahme weit mehr verworfen als mit niedrigeren Schwellenwerten, so dass auf weniger aussagekräftige Meta-Analysen (Random-Effects-Modelle) zurückgegriffen wird oder eine Meta-Analyse gar nicht in Frage kommt. Auch an dieser Stelle ist darauf hinzuweisen, dass eine inhaltliche Diskussion für den Umgang mit Heterogenität unerlässlich ist und nicht nur ein p-Wert mit obendrein sehr hoher Schwelle für Homogenität maßgeblich sein kann. Wir wären Ihnen für Erläuterungen dankbar, die diese Diskrepanz zum Cochrane Handbook erklären.

A.1.13 – Deutsches Netzwerk Versorgungsforschung e. V. (DNVF)

Deutsches Netzwerk
Versorgungsforschung e.V.

DNVF e.V. - Geschäftsstelle
c/o IMVR
Eupener Str. 129
50933 Köln

[redacted]

dnvf@uk-koeln.de

VR.Nr. 15170, Amtsgericht Köln

DNVF e.V.
c/o IMVR, Eupener Str. 129, 50933 Köln

AWMF
Frau Dr. Notacker
und
Herrn Prof. Selbmann

Stellungnahme des DNVF zur Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1 durch das IQWIG; Version vom 18.4.2013

Die aktuellen Änderungsvorschläge beziehen sich sehr explizit auf einige wenige Punkte der Version 4.0 vom 23.9.2011. Besonders relevant sind aus Sicht der Versorgungsforschung die Abschnitte zur endpunktbezogenen Bewertung (Abschnitt 3.1), sowie der Rationale zur Methodik zur Feststellung des Ausmaßes des Zusatznutzens (neuer Anhang).
Grundsätzlich begrüßen wir die Berücksichtigung von Lebensqualität als Zielgröße in der Nutzen-bewertung (Anhang, S. 28 und 33; Tabellen NT3, NT4, NT5) sowie den Plan, Verfahren multi-kriterieller Entscheidungsanalyse wie den Analytic Hierarchic Process und die conjoint Analyse anzuwenden (3.1.5, S. 13 – 14).
Allerdings erscheint der Umgang speziell mit dem Endpunkt Lebensqualität noch unzureichend methodisch dargelegt. So sind konkrete Standards für die Art, Auswahl und Auswertung der verwendeten Instrumente zu fordern. Von großer Wichtigkeit sind Kriterien für die Interpretation von Messwerten, insbesondere wenn ein Mix aus biomedizinischen und patientenberichteten End-punkten (PRO) zur Bewertung einer Maßnahme herangezogen wird. Es ist unklar, was die Autoren mit der Formulierung „zumindest unzweifelhaft etabliert“ meinen (S. 33). Sprachlich schwierig fassbar bzw. nicht nachvollziehbar ist eine Unterscheidung zwischen „beträchtlich“ und „erheblich“ (S. 33).
Zur Problematik von Schwellenwerten liegen bereits zur Version 4.0 umfangreiche und fundierte Stellungnahmen vor. Wir würdigen grundsätzlich die Notwendigkeit einer möglichst *a priori* festgelegten Entscheidungsgrundlage, um späterer Willkür in der Nutzenbewertung vorzubeugen. Allerdings darf nicht aus den Augen verloren werden, dass die jetzt getroffenen *a priori* Festlegungen ebenfalls willkürlich sind (die genannte Quelle ist kein empirischer Beleg für die Angemessenheit der angenommenen Schwellen). Immer ist die spezifische Versorgungskonstellation der jeweils unter-suchten Maßnahme mit zu berücksichtigen. Beispielsweise sind in einer hochpalliativen Situation oder bei fehlenden Therapiealternativen bei lebensbedrohlicher Erkrankung möglicherweise besondere Maßstäbe anzulegen. Das IQWIG sollte sich die Option offenhalten, gerade in ihrer Komplexität nicht vorhersehbaren Versorgungskontexten begründet von den festgelegten Kriterien abweichen zu können.

Bremen, 16. Mai 2013,
für den Vorstand des DNVF - Antje Timmer, Monika Klinkhammer-Schalke, Andrea Icks

A.1.14 – GKV-Spitzenverband

Spitzenverband

Stellungnahme des GKV-Spitzenverbands zum Papier des IQWiG: „Aktualisierung einiger Abschnitte der Allgemeinen Methoden 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden 4.1 (Entwurf vom 18. April 2014)“

Der GKV-Spitzenverband (GKV-SV) nimmt hiermit Stellung zu folgenden Abschnitten:

3.1.4 Endpunktbezogene Bewertung und 3.1.5 Zusammenfassende Bewertung (vorher Abschnitt 3.1.4)

3.3.3 Nutzenbewertung von Arzneimittel gemäß §35a SGBV
Neuer Anhang: Rationale der Methodik zur Festlegung

3.1.4 Endpunktbezogene Auswertung und 3.1.5 Zusammenfassende Bewertung (vorher Abschnitt 3.1.4):

Das IQWiG konkretisiert in 3.1.4 seine Anforderungen an die Beleglage für Aussagen zum Nutzen mit unterschiedlichen Aussagesicherheiten (Beleg, Hinweis, Anhaltspunkt, kein [Zusatz-]nutzen). Der bisherige Abschnitt 3.1.4 wird dazu in die o.g. Abschnitte unterteilt.

<u>Ableitung einer Nutzenaussage je Endpunkt (3.1.4):</u>

Das IQWiG differenziert - wie bisher- zwischen qualitativer Ergebnissicherheit (Studiendesign, mögl. Verzerrungen) und quantitativer Ergebnissicherheit (statist. Unsicherheit). Die Operationalisierung der Anforderungen an die Beleglage (s.Tabelle 2, S. 11) für unterschiedliche Aussagesicherheiten (Beleg, Hinweis, Anhaltspunkt) ist umfassender gestaltet worden, indem sie nunmehr die Bewertung der Effekte sowohl aus Einzelstudien als auch aus Meta-analysen darstellt und zwischen „deutlich gleichgerichteten“, „mäßig gleichgerichteten“ und „nicht gleichgerichteten“ Effekten in Abhängigkeit von der Anzahl der Studien differenziert. Ausgeführt wird zudem, in welchen Fällen eine Herabstufung der Ergebnissicherheit (Surrogatendpunkte, schwerwiegende Designmängel, Zweifel an die Übertragbarkeit auf die Behandlungssituation in D) bzw. Heraufstufung (u.a.große Effekte)

erfolgen kann. Der GKV-SV begrüßt die weitergehende Konkretisierung und Transparenz der Ableitung der Nutzenaussage je Endpunkt durch das IQWiG.

Ableitung endpunktübergreifender Bewertung (Abschnitt 3.1.5)

Das IQWiG stellt dar, dass es das endpunktübergreifende Fazit wie bisher in Form einer Nutzen-Schaden-Abwägung durch qualitative bzw. semiquantitative Abwägung der Effekte auf alle Endpunkte gegeneinander oder alternativ quantitativ durch einen endpunktgewichteten Summenscore zieht und dies themenspezifisch konkretisiert. Der GKV-SV begrüßt die kritischen Erwägungen des IQWiG zum QALY-Konzept und die Berücksichtigung alternativer Verfahren der Präferenzerhebung wie Analytic Hierachy Process (AHP) und Conjoint-Analyse (CA). Der GKV-SV spricht sich für eine weitere Konkretisierung aus, unter welchen Voraussetzungen das IQWiG AHP und CA im Rahmen des Fazits der Nutzen-Schadens-Abwägung in Betracht ziehen wird.

3.3.3 Nutzenbewertung von Arzneimitteln gemäß §35a SGB V und neuer Anhang „Rationale der Methodik zur Festlegung des Ausmaßes des Zusatznutzens“

In diesem Abschnitt und im neuen Anhang „Rationale der Methodik...“ integriert das IQWiG die Operationalisierung der Ableitung des Ausmaßes des Zusatznutzens aus dem Anhang A der Dossierbewertung zu Ticagrelor (A11-02, Seiten 86-92) in seine Methoden und stellt die Rationale dafür umfassender dar.

Der GKV-SV begrüßt grundsätzlich die Entwicklung eines Konzepts zur Ableitung des Ausmaßes des Zusatznutzens im Rahmen der durch das Gesetz und Rechtsverordnung vorgegebenen Kriterien der Nutzenbewertung nach 35a SGBV. Gleichwohl betrachtet es der GKV-SV als kritisch, dass die vom IQWiG vorgenommenen Ergänzungen der Kriterien gemäß AM-NutzenV und die Festlegung der quantitativen Schwellenwerte zur endpunktbezogenen Feststellung des Ausmaßes eines Effekts indikationsübergreifend und ohne inhaltlich begründete Ableitung auf Basis gesellschaftlicher Präferenzen erfolgt. Der GKV-SV begrüßt, dass das IQWiG bei der Ableitung der Gesamtaussage zum Ausmaß des Zusatznutzens die Problematik insofern adressiert, indem es ausführt, dass „eine strenge Formalisierung nicht

möglich [ist], da für die hierzu zu treffenden Werturteile gegenwärtig keine ausreichende Abstraktion bekannt ist".

Der GKV-SV begrüßt, dass das IQWiG nunmehr Transparenz schafft, wie es die Ableitung eines quantifizierbaren Ausmaßes nicht nur für binäre Zielgrößen (Relatives Risiko), sondern darauf basierend auch für andere Skalen von Zielgrößen operationalisiert.

Methodische Kommentare zum Anhang „Rationale der Methodik zur Festlegung des Ausmaßes des Zusatznutzens"

Im Anhang begründet das IQWiG ausführlich die Methodik zur Entwicklung der Schwellenwerte und dafür die Wahl relativer Effektmaße.

Schwellenwerte

Mit Hilfe der Einführung von neuen Schwellenwerten S, mit $S < 1$ (verschobene Hypothesengrenze), für die rechte Grenze der 95%-Konfidenzintervalle des Relativen Risikos wird ein Modell zur Klassifizierung des Zusatznutzens (in erheblich, beträchtlich, gering) formuliert.

Für die Berechnung der Schwellenwerte verwendet das IQWiG die Fallzahlformel für Pearsons χ^2-Test unter der Annahme des tatsächlichen Effekts (RR1): $KI = RR1\ (1 - 1 / \sqrt{2}) + 1/\sqrt{2}$. Diese Formel setzt allerdings die Verdoppelung der Fallzahlen voraus (um die gleiche Power zu erhalten). Dazu müssen entweder zwei gleichartige Studien (ähnliche Anzahl von Patienten, gleiche Power) zusammengefasst werden oder die Fallzahl in der betroffenen Studie erhöht werden.

Um die neuen Schwellenwerte zu berechnen, werden gewünschte Effekte definiert. So z.B. das Relative Risiko von 0,5 für die Zielgröße Gesamtüberleben. Daraus wird ein Schwellenwert für das 95%-KI von 0,85 berechnet. Diese gewünschten Effekte werden je nach Zusatznutzen und Zielgrößenkategorie unterschiedlich festgelegt. Die Schwellenwerte sollen je näher an 1 liegen (unterhalb von 1), je mehr Bedeutung der Zielgröße zugeordnet wird (und umgekehrt).

Dem GKV-SV ist die Motivation für die Festlegung der gewünschten Effekte ist nicht ersichtlich. Nur für den gewünschten tatsächlichen Effekt von RR = 0,5 wird auf die Quelle

Djulbegovic et al. 2008 verwiesen. In dieser Arbeit werde die Grenze 0,5 als „Durchbruch“ bezeichnet.

Im Fazit scheint die Festlegung der gewünschten tatsächlichen Effekte aus Sicht des GKV-SV als subjektiv und nicht ausreichend begründet. Auch ist für die Verwendung der obigen Formel eine weitere gleichartige Studie Voraussetzung, was auch nicht als selbstverständlich vorausgesetzt werden kann.

Wahl relativer Effektmaße

Das IQWiG versucht, eine Operationalisierung der Bewertung des Ausmaßes des Zusatznutzens vorzunehmen. In der Betrachtung wird auf relative Effektmaße abgestellt (z.B. RR, OR).

Mit Hilfe verschobener Hypothesengrenzen (<1) wird versucht, nach Endpunkten differenziert, eine Klassifizierung vorzunehmen. Dazu werden gewünschte tatsächliche Effekte angenommen. Diese Annahmen erscheinen willkürlich gewählt zu sein. Nur bei der angegebenen verschobenen Grenze (Endpunkt Gesamtmortalität) von 0,85 (gewünschter tatsächlicher Effekt RR=0,5) wird auf eine Arbeit von Djulbegovic et al. (2008) verwiesen.

Aus Sicht des GKV-SV ist die Berechnung der verschobenen Hypothesengrenzen (Schwellenwerte) überprüfenswert. Hier wird eigentlich zwingend vorausgesetzt, dass zwei gleichartige Studien (ähnliche Fallzahlen, gleiche Power, gleiche Zielvariablen, ähnliche Annahmen) zur Verfügung stehen. Wenn dies nicht der Fall ist, und mit derselben Anzahl von Patienten vorgegangen wird, verringert sich die Power.

Dies bedeutet z.B.: Beim Endpunkt Mortalität wird ein Zusatznutzen als erheblich eingestuft, wenn der rechte Randpunkt des 95%-KI kleiner als der Schwellenwert 0,85 ist, welcher mit der im Anhang angegebenen Formel aus dem gewünschten Effekt $RR_1=0,5$ errechnet wurde. Liegt jedoch nur eine Studie vor, die mit einer Power von z.B. 90% die Nullhypothese RR>=1 gegen RR<1 getestet hat, so wurde die Fallzahl n so geplant, dass bei Annahme des angestrebten Effekts $RR_1=0,5$ durch den Test mit 90% Wahrscheinlichkeit das Konfidenzintervall einen rechten Randpunkt hat, der kleiner als 1 (und nicht kleiner als 0,85) ist. Da die Fallzahl bei nicht vorhandener zweiter Studie unverändert bleibt, wird die Power des Tests zur verschobenen Hypothese RR>=0,85 gegen RR<0,85 wesentlich geringer als

90% sein. Das heißt, es sinkt die Wahrscheinlichkeit beträchtlich, trotz eines tatsächlichen Effekts von $RR_1=0,5$ ein Konfidenzintervall zu erhalten, dessen rechter Randpunkt kleiner als 0,85 ist. Anders ausgedrückt, es steigt die Wahrscheinlichkeit stark an, einen erheblichen Zusatznutzen (definiert durch $RR_1=0,5$) als höchstens beträchtlich einzuschätzen. Bei Existenz von nur einer Studie dient also das Verfahren höchstens als hinreichendes aber nicht als notwendiges Kriterium zur Einstufung als erheblicher Zusatznutzen. Analoge Überlegungen gelten für die Einstufung als beträchtlicher Zusatznutzen und andere Zielgrößen.

A.1.15 – GlaxoSmithKline GmbH & Co. KG (GSK)

Stellungnahme von GlaxoSmithKline (GSK) zum Entwurf des IQWiG „Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“ vom 18.04.2013

1. Stellungnahme zu Abschnitt 2.1.1 (Bericht) und Abschnitt 2.2.3 (Review der Produkte des Instituts)

GSK ist der Überzeugung, dass ein externes Review auch weiterhin zwingend erfolgen sollte und nicht – wie nun vom IQWiG vorgeschlagen – als optionaler Prozessschritt vorgesehen sein sollte. Ein externes Review dient der Qualitätssicherung der IQWiG Produkte und stellt somit sicher, dass externe methodische und fachliche Expertise in IQWiG Produkte adäquat und regelhaft einfließt.

2. Stellungnahme zu Abschnitt 3.1.4 (Endpunktbezogene Bewertung)

In diesem Abschnitt beschreibt das IQWiG die regelhaften Anforderungen an die Beleglage zur Ableitung von Aussagen mit unterschiedlichen Aussagesicherheiten. Darin werden in Tabelle 2 die neuen Anforderungen operationalisiert und im Vergleich zum bisherigen Methodenpapier die qualitative Ergebnissicherheit eingeführt und das Prädiktionsintervall als zusätzliches Maß zur Darstellung der Heterogenität benutzt. Dadurch werden nun vom IQWiG insgesamt 10 Fälle für die Aussagesicherheit operationalisiert - im Vergleich zu den 5 Fällen im bisherigen IQWiG-Methodenpapier)

Das in der Tabelle 2 zusammengefasste Konstrukt ist eine Eigen-Entwicklung des IQWiG und stellt nach Auffassung von GSK nicht den internationalen Stand der evidenzbasierten Medizin dar. In diesem Zusammenhang sei auf folgende Punkte hingewiesen:

Prädiktionsintervall

Für die Beurteilung der Heterogenität im Rahmen einer Meta-Analyse haben sich internationale Standards entwickelt. Zu diesen internationalen Standards zählt nach Einschätzung von GSK nicht das Prädiktionsintervall. Dies basiert auf der Beobachtung, dass das Prädiktionsintervall, von seltenen Ausnahmen abgesehen, in aktuell publizierte Meta-Analysen nicht benutzt wird. Auch die fehlende Implementierung in gängigen Statistik-Programmen unterstützt diese Einschätzung. Damit stellt sich zudem die Frage, wie die vom IQWiG geforderte Berechnung und Darstellung des Prädiktionsintervalls für das Nutzendossier umzusetzen ist.
Da nach Einschätzung des IQWiGs das Konzept des Prädiktionsintervalls bei mindestens 4 vorliegenden Studien anzuwenden ist, stellt sich außerdem die Frage, ob Nutzendossiers mit mehr als 4 Studien im Vergleich zu Nutzendossiers mit 4 oder weniger Studien durch den beschriebenen Algorithmus gleich behandelt werden.

Schwellenwerte

Vom IQWiG werden eine Reihe von „Schwellenwerten“ etabliert (z.B. „Das Gesamtgewicht dieser Studien ist >= **80%**“, „Mindestens **2** dieser Studien zeigen statistisch signifikante Ergenisse“, „Mindestens **50%** des Gewichts dieser Studien basiert auf statistisch signifikanten Ergebnissen“ – Hervorhebung der Schwellenwerte durch GSK). Es fehlt eine Rationale für die Wahl dieser Schwellenwerte

Beleg mit einer Studie

Es wird zwar (im Text auf Seite 10) auf den Ausnahmefall verwiesen, dass auch mit einer Studie ein Nutzenbeleg abgeleitet werden kann – allerdings fehlt dieser Sachverhalt in Tabelle 2.

Berücksichtigung aller verfügbaren Evidenz

Unterhalb von Tabelle 2 heißt es:
„Liegen mehrere Studien mit unterschiedlicher qualitativer Ergebnissicherheit vor, so werden zunächst nur Studien mit der höherwertigen Ergebnissicherheit betrachtet und auf dieser Grundlage Aussagen zur Beleglage gemäß Tabelle 2 abgeleitet".
Es ist für GSK nicht nachvollziehbar, wieso grundsätzlich nicht alle verfügbare Evidenz für die Beurteilung der Ergebnissicherheit herangezogen werden soll.

Ausnahmen von der regelhaften Operationalisierung:

Es wird dargelegt, dass neben der regelhaften Operationalisierung – wie in Tabelle 2 des entsprechenden Abschnitts dargestellt – in begründeten Fällen weitere Faktoren diese Einschätzungen beeinflussen können. Konkret wird zu diesen weiteren Faktoren aufgeführt:

„Die Betrachtung von Surrogatendpunkten (siehe Abschnitt 3.1.2), das Vorliegen schwerwiegender Designmängel bei einer Studie, oder auch begründete Zweifel an der Übertragbarkeit auf die Behandlungssituation in Deutschland können z. B. zu einer Verringerung der Aussagesicherheit führen. Auf der anderen Seite können z. B. große Effekte oder eine eindeutige Richtung eines vorhandenen Verzerrungspotenzials eine Erhöhung der Sicherheit begründen."

Aus Sicht von GlaxoSmithKline ist das beschriebene Vorgehen bzgl. nachfolgender Punkte zu hinterfragen.

1) Im Gegensatz zum Bestreben des IQWiG eine Vergleichbarkeit der Nutzenbewertungen sicherzustellen, führt die nicht abschließende Definition von „begründeten Ausnahmefällen" zu einem Entscheidungsspielraum auf Seiten des IQWiG, der grundsätzlich unterschiedliche Einschätzungen zu an sich vergleichbarer Evidenz ermöglicht.

2) Aus Sicht von GlaxoSmithKline ist es gänzlich nicht nachvollziehbar inwiefern die alleinige Betrachtung von Surrogatparametern in Studien die Aussagesicherheit verringern kann. Aus der bloßen Betrachtung von Surrogatparametern lässt sich kein Einfluss auf die, in der Nutzenbewertung zu berücksichtigenden, patientenrelevanten Endpunkte ableiten. Diesbezüglich sollte primär auf die zugrunde liegende statistischen Analyse und die Fallzahl zurückgegriffen werden, um die Ergebnissicherheit beurteilen zu können.

3) GlaxoSmithKline stimmt dem IQWiG grundsätzlich zu, dass die Übertragbarkeit von Studienergebnissen auf die Behandlungssituation in Deutschland zu berücksichtigen ist. Allerdings folgt GSK nicht der Logik des IQWiG, dass dies im Rahmen der Beurteilung der Ergebnissicherheit zu erfolgen hat. Dies ist insbesondere deshalb der Fall, da vom pharmazeutischen Unternehmer im Nutzendossier spezifische Subgruppenanalysen vorzulegen sind, in denen unter anderem der Einfluß von Zentrums- und Ländereffekten untersucht wird. Die Übertragbarkeit der Studienergebnisse lässt sich daher primär anhand dieser Subgruppenanalysen ableiten. Das vorgeschlagene Vorgehen des IQWiG führt somit zu einer unnötigen Doppelbetrachtung der Übertragbarkeit von Studienergebnissen, die der Gemeinsame Bundesausschuss bereits – in methodisch eindeutig definierten Weise – im Rahmen von Subgruppenanalysen vornimmt.

3. Stellungnahme zu Abschnitt 3.3.3 (Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V)

<u>Operationalisierung zur Feststellung des Ausmaß des Zusatznutzens</u>

Es wird vom IQWiG spezifisch auf die Operationalisierung der Feststellung des Ausmaßes des Zusatznutzens eingegangen und insbesondere entsprechende Schwellenwerte definiert, um eine quantitative Aussage über das Ausmaß der Effektstärke festzustellen.
Das IQWiG beschreibt ferner im Anhang des Aktualisierungsvorschlags die Rationale für die beschriebene Methodik. Hierbei betont das IQWiG selbst, dass diesbezüglich notwendigerweise zu treffende Werturteile möglichst gering zu halten und explizit zu machen sind. Aus Sicht des IQWiG bedarf es daher einer

- expliziten Operationalisierung, um ein transparentes und nachvollziehbares Verfahren sicherzustellen, sowie einer
- abstrakten Operationalisierung, um größtmögliche Konsistenz zwischen den Nutzenbewertungen zu erzielen.

GSK stimmt dem IQWiG grundsätzlich zu, dass die Operationalisierung durch ein transparentes und nachvollziehbares Verfahren sicherzustellen ist. Allerdings ist die Methodik des IQWiG insbesondere aufgrund nachfolgender Punkte zu hinterfragen:
Die Basis für die Etablierung einer der in Tabelle NT1 dargestellten Schwellenwerte stellt eine Übersichtsarbeit aus der Onkologie (Djubelgovic 2008) dar. Dazu drängen sich zwei Fragen auf, die im vorliegenden Entwurf nicht adäquat adressiert werden: Gibt es weitere Arbeiten, die die dort dargelegte Evidenz unterstützen? Mit welcher Begründung kann eine Arbeit aus der Onkologie als Basis für die Herleitung von Schwellenwerten für alle Indiaktionsgebiete benutzt werden?

Die Herleitung der Schwellenwerte geht ferner vom „Regelfall des Vorliegens von zwei (z.B. pivotalen) Studien“ mit identischer Fallzahl aus. Dieser „Regelfall“ trifft jedoch in vielen Zulassungssituationen nicht zu.

Die vom IQWiG für die weiteren der in Tabelle NT1 dargestellten Schwellenwerte vorgenommene „Rasterung von 1/6 für die tatsächlichen Effekte“ wird vom IQWiG als „pragmatische Lösung“ bezeichnet. Es fehlt allerding eine wissenschaftlich fundierte Rationale für diese willkürlich erscheinende Rasterung. Zumindest eine wissenschaftliche oder sogar gesellschaftliche Diskussion der zugrunde liegenden Schwellenwerte erscheint erforderlich.

Die vom IQWiG abgeleiteten Schwellenwerte der Tabelle NT1 haben ihren Ausgangspunkt in einem relativen Risiko. Das IQWiG schlägt vor bei stetigen Zielgrößen auf Responderanalysen zurückzugreifen und somit eine Dichotomisierung eines stetigen Parameters vorzunehmen. Damit geht bekanntermaßen ein großer Informationsverlust einher, der nicht im Einklang mit den internationalen Standards der evidenzbasierten Medizin ist. Dieser wesentliche Mangel ist ein weiterer Beleg dafür, dass die vom IQWiG entwickelte Methodik als noch nicht ausgereift angesehen werden kann.

Zusammenfassend trägt eine indikationsübergreifende Betrachtung auf Basis von willkürlich abgeleiteten Schwellenwerten aus der Onkologie nicht der Heterogenität unterschiedlicher Indikationen und Schweregrade Rechnung. Im Hinblick auf die Vergleichbarkeit von Nutzenbewertungen scheint es aus Sicht von GSK ausreichend, sich auf eine Vorgehensweise zu verständigen, anhand derer eine indikations- und interventionsspezifische Operationalisierung im Rahmen einer Nutzenbewertung erfolgt.
Im Gegensatz zum Vorschlag des IQWiG würde GSK einen Indikations- und Endpunkt- bzw. Instrumentspezifischen Ansatz befürworten.

<u>Umgang mit ausreichend validen Surrogaten:</u>

Von Seiten GlaxoSmithKline ist es sehr zu begrüßen, dass das IQWiG ausdrücklich die Situation aufgreift, in der „ein ausreichend valides Surrogat“ aber keine formale Validierung vorliegt. In Hinblick hierauf geht das IQWiG einen pragmatischen Weg, der die Realität im Versorgungsalltag über methodische Anforderungen stellt, die teilweise schlicht aufgrund der zur Verfügung stehenden Datenlage nicht zu erfüllen sind. Ferner ist zu begrüßen, dass das IQWiG auf eine umfassende Definition des Begriffs „ausreichend valides Surrogat“ eingeht. Aus Sicht von GSK ist dies insbesondere erforderlich, da keine abschließende, indikationsübergreifende Definition abzuleiten ist. Gleichermaßen sollte dieser Ansatz auch bei der Operationalisierung der Feststellung des Ausmaßes des Zusatznutzens erfolgen, wobei das IQWiG die Heterogenität von Erkrankungen und Behandlungen ausblendet und einen „one size fits all“-Ansatz verfolgt.

A.1.16 – Herescon GmbH

herescon gmbh
health economic research & consulting

IQWiG
Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1. Entwurf vom 18.04.2013
Stellungnahme der Herescon GmbH
22.05.2013

Herescon GmbH, Königsworther Straße 2, 30167 Hannover

Kontakt:
Dr. Werner Kulp
Telefon: +49 511 897 093 10
Fax: +49 511 897 038 48
info@herescon.com

Die nachfolgende Stellungnahme der Herescon GmbH bezieht sich auf die vom IQWiG vorgelegte Aktualisierung einiger Abschnitte der Allgemeinen Methoden 4.0 sowie zum neuen Abschnitt zur Erstellung der Allgemeinen Methoden Version 4.1 im Entwurf vom 18.04.2013.

Grundsätzlich begrüßen wir die Aktualisierungen bzw. neuen Abschnitte. Es ergeben sich aus unserer Sicht jedoch für die Kapitel 3.1.4, 3.1.5 sowie 3.3.3 folgende Punkte mit fehlendem Detaillierungsgrad, welche nachfolgend mit der Bitte um Klärung, Nachbesserung bzw. Konkretisierung dargestellt werden:

3.1.4 Endpunktbezogene Bewertung

S. 9: „Das Institut verwendet die folgenden drei Kategorien zur Graduierung des Ausmaßes der qualitativen Ergebnissicherheit auf Einzelstudien- und Endpunktebene:

- *hohe qualitative Ergebnissicherheit: Ergebnis einer randomisierten Studie mit niedrigem Verzerrungspotenzial.*
- *mäßige qualitative Ergebnissicherheit: Ergebnis einer randomisierten Studie mit hohem Verzerrungspotenzial.*
- *geringe qualitative Ergebnissicherheit: Ergebnis einer nicht randomisiert vergleichenden Studie. […]“*

Stellungnahme:

- Die vom IQWIG vertretene Position, eine nicht-randomisiert vergleichende Studie habe, auch wenn methodisch sorgfältig durchgeführt, stets eine niedrigere qualitative Ergebnissicherheit als eine randomisierte Studie mit hohem Verzerrungspotential erscheint insbsd. in Bezug auf Beobachtungsstudien zu unerwünschten Arzneimittelwirkungen problematisch. Nach der Marktzulassung stellen methodisch valide durchgeführte Beobachtungsstudien (pharmakoepidemiologische Studien) eine wichtige Quelle für die Identifizierung und Quantifizierung von Arzneimittelrisiken dar. Nach der vorliegenden IQWIG-Methodik wären solchen Studien aber generell – allein aufgrund ihres Studiendesigns – noch unter dem Stellenwert eines stark verzerrten RCTs einzuordnen. Dieses wird der Bedeutung derartiger Studien für die Identifizierung und Quantifizierung von Arzneimittelrisiken nicht gerecht und bedarf der Überarbeitung.

- Es bleibt unklar, wann das IQWIG auch nicht-randomisierte vergleichende Studien (mit aus Sicht des IQWIG geringer qualitativer Ergebnissicherheit) bei der Bewertung berücksichtigen will und wann die Bewertung auf RCTs beschränkt werden soll. Bei der Nutzenbewertung von bereits zugelassenen Arzneimitteln, für die u.U. schon post-marketing Beobachtungsstudien zu Nutzen- und Schadenaspekten existieren, sollte eine Berücksichtigung dieser Studien bei der Nutzenbewertung vorgesehen werden. Andernfalls könnten durch die Beschränkung der Einschlusskriterien und der systematischen Literatursuche auf RCTs wesentliche Bewertungsaspekte (insbesondere in Bezug auf Schaden) aus der Bewertung ausgeschlossen werden.

S. 10: „[…] Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und Ihre Ergebnisse besondere Anforderungen zu stellen. […]"

Stellungnahme:

- Die Formulierung *„[…] besondere Anforderungen […]"* bedarf einer formal inhaltlichen Konkretisierung. Es wird nicht klar, inwiefern sich dieses Anforderungsprofil auf eine *„hohe qualitative Ergebnissicherheit"* (RCT mit niedrigem Verzerrungspotential) stützt.

3.1.5 Zusammenfassende Bewertung

S. 13: „Eine Möglichkeit der gleichzeitigen Würdigung von Nutzen und Schaden ist die Gegenüberstellung der endpunktbezogenen Nutzen- und Schadenaspekte. Dabei werden die Effekte auf alle Endpunkte (qualitativ oder semiquantitativ) gegeneinander abgewogen mit dem Ziel, zu einer endpunktübergreifenden Aussage zum Nutzen bzw. Zusatznutzen einer Intervention zu kommen. Eine weitere Möglichkeit der gleichzeitigen Würdigung besteht darin, die verschiedenen patientenrelevanten Endpunkte zu einem einzigen Maß zu aggregieren. […]"

Stellungnahme:

- Da das IQWiG die Aggregierung zu einem einzigen Summenscore im Berichtsplan oder im Vorbericht bewerten möchte, sind aus methodischen und verfahrensbezogenen Gründen Angaben zu jenen Kriterien erforderlich, anhand derer eine solche Bewertung durchgeführt wird.
- Es sollte ferner dargelegt werden, wie sich eine *„semiquantitative"* Abwägung ausgestaltet.
- Es sollte diskutiert werden, inwieweit weitere Parameter des zu bewertenden Wirkstoffes (z.B. Indikation, therapeutischer Stellenwert im Indikationsgebiet), ggf. unter Einbeziehung medizinischer (oder anderweitiger indikationsspezifischer) Expertise Berücksichtigung finden sollten.

S. 13: „Häufig werden sogenannten Nutzwerte für Gesundheitszustände […] Aufgrund der ethischen und methodischen Probleme gerade der häufig verwendeten QALYs sollten alternative Verfahren der multikriteriellen Entscheidungsfindung oder Präferenzerhebung angewendet werden. Dazu zählen u.a. der Analytical Hierachy Process (AHP) und die Conjoint-Analyse (CA)."

Stellungnahme:

- Die QALY-Kritik ist methodologisch berechtigt, aber auch nicht neu; doch liefert kein anderes Konzept international validierte Werte für die einheitliche Bewertung von Gesundheitsleistungen.
- Eine kritische Diskussion des QALY-Konzeptes ist letztlich sachlich richtig, in Relation zu den Ausführungen zum AHP und zur CA allerdings wenig wissenschaftlich ausgewogen. Insbesondere sollte dargestellt werden, inwieweit die genannten Instrumente auf den Frühbewertungskontext insb. zur Bewertung und Quantifizierung des Zusatznutzens übertragen werden können.
- Diese Verfahrensalternativen können aus unserer Sicht als Ergänzung zum QALY-Konzept gesehen werden.
- Es erscheint fraglich, ob es zielführend ist, ein etabliertes mikroökonomisches Konzept mit bekannten und beschriebenen Defiziten durch andere Konzepte zu ersetzen, die ebenso von Werturteilen abhängig sind und deren Einschränkungen sowie deren Praktikabilität nicht umfänglich bekannt sind. Bekannte methodische Defizite sind kein hinreichender Grund eine Methodik gänzlich zu verwerfen.

S. 14: „[…] Die am AHP Teilnehmenden werden dann jeweils binär zu den Kriterien befragt, d.h. sie müssen auf einer vorgegebenen Skala wählen, wie viel mehr ihnen ein Kriterium als ein anderes Kriterium bedeutet."

Stellungnahme:

- Es sollte dargestellt werden, wie sich die Gruppe der *„Teilnehmenden"* zusammensetzt (Patienten, Versicherte). Da diese Wahl nicht frei von Werturteilen ist, ist diese zumindest explizit zu begründen.
- Fraglich ist zudem, ob die Festlegung dieses Personenkreises innerhalb des Mandatsbereiches des IQWiG liegt oder eine Normkonkretisierung durch den G-BA bedarf.

3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V

Stellungnahme zum Gesamtkapitel:

Das IQWIG legt in seiner Überarbeitung der Methoden einen Ansatz vor, um das Ausmaß des Zusatznutzens anhand von relativen Effektmaßen zu definieren. Dieser Ansatz ist vom IQWIG erstmals bei der Nutzenbewertung nach AMNOG zur Substanz Ticagrelor verwendet worden. Es ist nachvollziehbar und begrüßenswert, dass das IQWIG eine transparente Methodik zur Bewertung des Ausmaßes des Zusatznutzens vorlegen will, der vorgeschlagen Ansatz erscheint aber aus folgenden Gründen als Entscheidungsgrundlage für eine Nutzenbewertung noch nicht ausgereift:

- Es handelt sich um keine etablierte Methodik zur Einschätzung des Ausmaßes eines Zusatznutzens. Es liegen keine wissenschaftlichen Arbeiten vor, die die Geeignetheit des Ansatzes für die Nutzenbewertung untersucht haben. Das IQWIG hat diesen methodischen Ansatz erstmals bei der Nutzenbewertung zum Wirkstoff Ticagrelor eingesetzt (Datum: 29.09.2011). Es erscheint unverständlich, warum das Institut in der Zwischenzeit trotz der von Fachgruppen vorgebrachten Kritik (u.a. AkdÄ; VFA) an diesem Ansatz keine wissenschaftlichen Untersuchungen zur Geeignetheit (Plausibilität, Validität, Reliabilität) des vorgeschlagenen Verfahrens durchgeführt hat bzw. ausgeschrieben hat. Ein Ansatz, eine solche Methodik hinsichtlich ihrer Geeignetheit zu überprüfen könnte z.B. die Anwendung des Instrumentes auf eine

ausreichend große Anzahl von „Beispielfällen“ der Arzneimittel-Nutzenbewertung durch verschiedene Experten sein, die dann mit dem in einem Konsensverfahren gewonnenen Allgemeineinschätzung zum jeweiligen Zusatznutzen (quasi als Goldstandard) abgeglichen werden könnte.

- Es ist nicht ersichtlich, ob das Institut vor Festlegung auf die vorgeschlagene Methodik eine systematische Evaluation aller verfügbaren (publizierten) wissenschaftlichen Ansätze zur Quantifizierung von Nutzen und Schaden bzw. der Verhältnismäßigkeit von Nutzen und Schaden (Benefit-Risk-Assessment) vorgenommen und auf ihre Geeignetheit für die Nutzenbewertung in Deutschland überprüft hat. Angesichts der Bedeutung der Methodik für das Ergebnis der Nutzenbewertung (und nachfolgende Preisverhandlungen) erscheint ein derartiges, transparentes Vorgehen dringend erforderlich.
- Sollte die Entscheidung für ein neues, bislang nicht publiziertes Verfahren fallen (wie z.B. das jetzt vorgeschlagene Vorgehen), so ist zusätzlich zur Einarbeitung in das IQWIG-Methodenpapier eine (vorzugsweise internationale) wissenschaftliche Publikation dieser Methodik zu fordern, damit auch der (internationalen) Fachöffentlichkeit die Möglichkeit zur Diskussion gegeben wird. Die Veröffentlichung auf der Homepage des IQWIG erscheint hingegen wissenschaftlich unzureichend.

S. 19: „Zur Feststellung des Ausmaßes des Zusatznutzens bei stetigen oder quasi stetigen Zielgrößen werden Responderanalysen herangezogen. Dazu bedarf es eines validen bzw. etablierten Responsekriteriums bzw. Cut-off-Wertes. […]“

Stellungnahme:

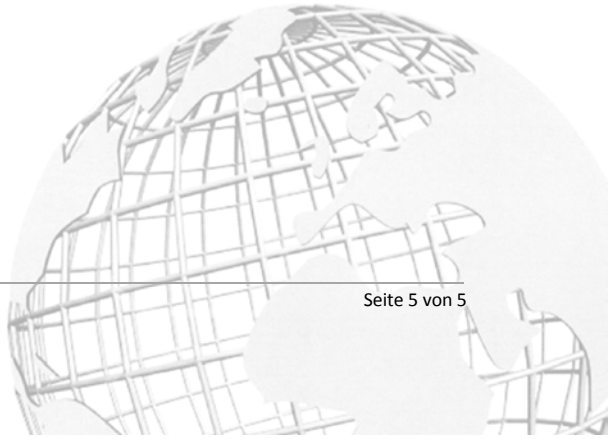
- Problematisch ist, dass es die vom IQWiG geforderten validierten Responseschwellenwerte vielfach nicht gibt. Es sollte dargelegt werden, wie im Fall fehlender Schwellenwerte kardinale Skalenwerte dichotomisiert werden sollen.
- Die vorgeschlagene Lösung zu Feststellung des Ausmaßes des Zusatznutzens bei stetigen oder quasi-stetigen Zielgrößen ist unbefriedigend, da sie die Analyse mittels Responderanalyse zwingend voraussetzt. Sind derartige Analysen nicht vorhanden und auch durch den PU nicht durchführbar, da kein Zugriff auf die Originaldaten besteht (z.B. bei der Durchführung von indirekten Vergleichen), lassen sich diese Daten nicht nutzen (obgleich sie als Teil der vorhandenen Evidenz zu berücksichtigen wären). Es sollte eine Methode vorgelegt werden, die auch die Einbeziehung stetiger Zielgrößen ermöglicht ohne eine Responderanalyse vorauszusetzen.

Anhang zum Abschnitt 3.3.3: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens

S. 32/33: „Dazu wurde ursprünglich ein relatives Risiko von 0,50 – […] - als Effekt erheblichen Ausmaßes für die Zielgröße Gesamtmortalität verankert. […] Sie führten dazu, dass für einen Schwellenwert von 0,85 die gleichzeitige Anforderung nach Realisierbarkeit und Stringenz als erfüllt angesehen werden kann. […] Eine Rasterung von 1/6 für die tatsächlichen Effekte erwies sich als pragmatische Lösung.“

Stellungnahme:

- Wie schon mehrfach erwähnt, stützt sich dieses Konzept auf eine Publikation ausschließlich aus dem onkologischen Indikationsbereich. Mit Blick auf andere Indikationen ist es fraglich ob ein RR von 0,50 als ausreichend angesehen werden kann.
- Ferner ist unklar, ob die vom IQWiG pragmatisch (oder auch arbiträr) abgeleiteten Schwellenwerte indikations- und populationsübergreifend anwendbar sind. Es ist fraglich, dass verschiedene Endpunkte in unterschiedlichen Indikationen denselben Raum für therapeutische Verbesserungen halten.



A.1.17 – Janssen-Cilag GmbH

Stellungnahme der Janssen-Cilag GmbH, Neuss, zum Entwurf der Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1

22.05.2013

Die Janssen-Cilag GmbH bezieht zu den folgenden Punkten (in der Reihenfolge ihres Auftretens im vom IQWiG vorgelegten Entwurf) Stellung:

S. 13: Zusammenfassende Bewertung und Nutzung von Aggregationsmethoden

Das IQWIG diskutiert auf Seite 13 ff. die Möglichkeit einer zusammenfassenden Bewertung der einzeln bewerteten Endpunkte im Rahmen einer Nutzenbewertung. Um zu einer aggregierten Bewertung zu kommen, können die verschiedenen Endpunkte zu einem einheitlichen Maß aggregiert werden, wobei die einzelnen Endpunkte mit Gewichten versehen werden, welche deren Wichtigkeit widerspiegeln. Methodisch kann dies mit einer Erhebung von Nutzwerten gemacht werden, die im Rahmen des QALY Ansatzes verwendet werden. Alternativ gibt es Verfahren der multikriteriellen Entscheidungsfindung oder der Präferenzerhebung wie der Analytic Hierarchy Process (AHP) oder die Conjoint-Analyse (CA), bei der jeweils Patienten befragt werden.

Hierzu ist anzumerken:

- Nutzung im Rahmen der frühen Nutzenbewertung: Die genannten Verfahren werden im Kapitel zur Nutzenbewertung thematisiert, nicht aber im Abschnitt zur frühen Nutzenbewertung nach § 35a SGB V. Die frühe Nutzenbewertung steht allerdings vor der gleichen Problemstruktur: Endpunkte müssen in ihrer Wichtigkeit eingeschätzt und zu einem Gesamturteil aggregiert werden (im Rahmen der frühen Nutzenbewertung zu einer einzigen Einschätzung des Zusatznutzens für ein neues Medikament bzw. für eine Teilpopulation; vgl. dazu die auf S. 16 erwähnte Gesamtschau). Es ist bisher nicht klar – weder in den Methoden des IQWiG noch in der Vorgehensweise des Gemeinsamen Bundesausschuss – wie genau diese Aggregation im Rahmen der frühen Nutzenbewertung erfolgt. Die oben genannten Methoden (AHP, CA) bieten die Möglichkeit, dies systematisch und auf das Urteil einer größeren Zahl von Patienten gestützt zu tun. Insofern ist anzuregen, dass aus diesem Grund sowie aus Gründen der Konsistenz der Vorgehensweise zwischen früher Nutzenbewertung und Nutzenbewertung die genannten Verfahren auch in die Endpunktgewichtung im Rahmen der frühen Nutzenbewertung einfließen. Die entsprechenden Abschnitte S. 13f. können hierfür in die Abschnitte zur frühen Nutzenbewertung übernommen werden. Grundsätzlich sollte dabei darauf hingewiesen werden, dass die genannten Methoden einer weiteren Erforschung und insofern ständigen Überprüfung ihrer Eignung für den Kontext einer Nutzenbewertung bedürfen.

- Prospektive Planung: Das IQWIG sieht bei der Verwendung der Gewichtungsmethoden eine prospektive Festlegung der Gewichtung vor (S. 13). Dies ist nicht zwingend notwendig und teilweise nur schwer umzusetzen, da eine Conjoint Analyse in der Regel zweistufig erfolgt. In einem ersten Schritt wird die Wichtigkeit aller Endpunkte anhand einer Analogskala erfasst.

Die abgefragten Endpunkte werden anschließend mit Hilfe einer Faktoranalyse reduziert, und den Patienten wird dann im Rahmen einer CA diese reduzierte Anzahl an Endpunkten vorgelegt. Die Reduktion der Anzahl der Endpunkte ist notwendig, weil die Probanden mit der Bewertung einer hohen Anzahl an Endpunkten überfordert wären und dadurch die Validität der CA leiden würde. Ein Beispiel: Eine aktuelle Patientenpräferenzstudie im HIV Bereich hat beispielsweise 26 abgefragte Therapieeigenschaften faktoranalytisch zu sechs Stimuli für die CA reduziert (Mühlbacher et al. 2013). Da aber a priori nicht klar ist, welche Endpunkte zusammengefasst werden – es handelt sich hier um eine empirische Beobachtung – können im Vorfeld auch nicht alle Endpunkte festgelegt werden, die letztendlich in die CA eingehen. Die prospektive Planung der CA bereits im Rahmen der klinischen Studien ist daher nicht notwendig. Auch im Nachhinein ist eine verzerrungsfreie Endpunktgewichtigung möglich. Der entsprechende Hinweis auf S. 13 ist daher zu streichen.

S. 16: Nutzenkategorien in der frühen Nutzenbewertung

Das IQWiG setzt sich im Abschnitt 3.3.3 mit der Nutzenbewertung nach § 35a SGB V auseinander und dort unter anderem auch mit der Kategorie des „nicht quantifizierbaren Zusatznutzens". Nicht quantifizierbar sei nach Auffassung des IQWiG unter anderem dann zu vergeben, wenn ein Effekt auf ein ausreichend valides Surrogat nachgewiesen sei, eine „verlässliche Schätzung" für den jeweiligen Effekt der patientenrelevanten Zielgröße nicht möglich sei. In dieser Weise argumentiert das IQWiG bei frühen Nutzenbewertungen unter anderem im Falle der Parameter nachhaltiges virologisches Ansprechen (SVR) bei Hepatitis C (IQWiG 2012a) sowie virologisches Ansprechen bei HIV (IQWiG 2012b), wobei jeweils Mortalitäts- oder Morbiditätsendpunkte (z.B. Vermeidung von hepatozellulärer Karzinome im Falle von Hepatitis C) als final patientenrelevant angesehen werden.

Entscheidend bei den Ausführungen im Methodenpapier ist, dass IQWiG darauf verweist, dass eine „verlässliche Schätzung" (S. 16) des patientenrelevanten Effekts nicht möglich sei. Nähere Ausführungen dazu, was verlässlich ist oder nicht verlässlich werden im Methodenpapier nicht gemacht. Allerdings wäre genau dies angesichts der besonderen Fallkonstellation einer frühen Nutzenbewertung notwendig; es sollte angegeben und diskutiert werden, welches Maß an Verlässlichkeit ausreicht, um einen patientenrelevanten Effekt in obiger Fallkonstellation quantifizieren zu können.

Ein gutes Beispiel hierfür ist der Parameter nachhaltiges virologisches Ansprechen in der Hepatitis C (SVR) – auch wenn der Gemeinsame Bundeausschuss hier am Ende dem IQWiG nicht gefolgt ist und den Parameter am Ende für eindeutig patientenrelevant erklärt hat (die ausführliche Diskussion im Rahmen der frühen Nutzenbewertung soll hier entsprechend nicht wiederholt werden; vgl. dazu die Dokumentation Gemeinsamer Bundesausschuss 2012b und die dortigen Stellungnahmen). Aufgrund der Natur der Hepatitis C-Erkrankung (langsam fortschreitend über 20-30 Jahre) ist die Ermittlung des Effekts von neuen Behandlungsmethoden auf Endpunkte wie Mortalität oder die Entwicklung von Karzinomen in der Regel nicht möglich; Studien nutzen daher üblicherweise SVR (also die Virusfreiheit 24 Wochen nach Ende der Behandlung) als Endpunkt; dieser ist weltweit in der Fachwelt anerkannt. Auch die Abschätzung dieses Effekts in Validierungsstudien ist praktisch kaum machbar, da diese zu lange und ethisch nicht vertretbar wären. Allerdings ist Evidenz aus observationellen Studien verfügbar, die sehr eindeutig auf einen deutlichen Effekt hinweist. In solchen Fällen sollte eine Quantifizierung durchgeführt werden und nicht auf die mangelnde Verlässlichkeit der Zusammenhänge verwiesen werden. Der Grad der Verlässlichkeit, der für eine Aussage ausreicht, bedarf in jedem Fall einer genaueren wissenschaftlichen Diskussion.

Ebenso sei darauf verwiesen, dass solche Fälle (ausreichend valides Surrogat vorhanden, Effekt auf weitere Endpunkte unklar) auch ein Einsatzpunkt für Modellierungen des Krankheitsverlaufs sind. Im Falle der oben erwähnten Hepatitis C-Bewertungen wurde auf entsprechende Ansätze verwiesen (vgl. dazu Gemeinsamer Bundesausschuss 2012b). Die Modellierung von Krankheitsverläufen ist auch in anderen Therapiefeldern eine anerkannte Methode, um bei Nichtdurchführbarkeit entsprechender Studien den Effekt von Therapien oder Therapiesequenzen zu ermitteln. Ein Beispiel ist unter anderem die Ermittlung optimaler Behandlungssequenzen in der Onkologie bei Vorhandensein sehr vieler Therapiemöglichkeiten. Basierend auf vorhandenen Studiendaten können deren Ergebnisse derart synthetisiert werden und mittels international anerkannter Entscheidungsmodelle (z.B. Markov) als prognostische Modellstudie derart dargestellt werden, dass das Ergebnis einer klinisch nicht umsetzbaren Studie, mit einer hohen Wahrscheinlichkeit nachgestellt und prognostiziert wird (vgl. als Beispiel Heeg et al. 2010). Derartige Modellierungsansätze lassen sich auch für eine Quantifizierung des Zusatznutzens in der vom IQWiG beschriebenen Konstellation nutzen und sollten entsprechend in das Methodenpapier einfließen.

Zudem sei darauf verwiesen, dass der Gemeinsame Bundesausschuss weitere Fallkonstellationen sieht, bei denen ein nicht quantifizierbarer Zusatznutzen zum Tragen kommen kann. Verwiesen sei auf ein Expertengespräch vom 22. März 2012, in dem vom damaligen Vorsitzenden des Gemeinsamen Bundesausschuss geäußert wurde: „Es sei durchaus möglich, dass ein Zusammenführen der Gruppen mit unterschiedlichem Zusatznutzen-Ausmaß zum Teil zu einem nicht quantifizierbaren Zusatznutzen führe.“ (Gemeinsamer Bundesausschuss 2012a) Der Klarheit wegen sollte aufgeführt werden, in welcher Relation diese Fallkonstellation zu der vom IQWiG angegebenen Konstellationen steht.

S. 26 ff: Methodik zur Feststellung des Zusatznutzens

Das Verfahren, mit dem das IQWiG das Ausmaß des Zusatznutzens bestimmt, stellt keinen internationalen Standard der evidenzbasierten Medzin dar und ist ein Eigenkonstrukt des IQWiG. Die Quantifizierung der Feststellung des Ausmaßes des Zusatznutzens basiert auf den Schwellenwerten für die obere Grenze des 95%-Konfidenzintervalls und wird über das tatsächliche Relative Risiko auf Basis von Djulbegovic 2008 abgeleitet. Die Ableitung des Zusatznutzens basierend alleinig auf dieser Quelle ist somit eine „willkürlich“ getroffene Entscheidung, zumal sie sich auch nicht auf allgemein anerkannte Schwellenwerte beruft. Außerdem hat Djulbegovic seinen Vorschlag mit dem Hintergrund der Onkologie entwickelt, also einer ganz spezifischen Indikation; das IQWiG macht aber keinen indikationsspezifischen Vorschlag sondern präsentiert ein indikationsunabhängiges starres Verfahren, in dem alle Indikationen gleich bewertet werden. Es stellt sich auch die Frage, ob ein schematisches ex-post Vorgehen in diesem Zusammenhang zulässig oder sinnvoll ist.

Ein weiteres Problem stellt sich dadurch, dass die Einordnung in die Kategorien (Symptome, Nebenwirkung, Lebensqualität) nicht eindeutig ist. So werden beispielsweise Symptome und Nebenwirkungen in einer Kategorie zusammengefasst. Auch wurde die Frage nicht diskutiert, ob die Liste der Zielgrößenkategorien vollständig ist und die Intentionen des Gesetzgebers vollständig abdeckt. Es findet auch kein Einbezug von Patientenpräferenzen oder Präferenzen der Gesamtbevölkerung, bzw. eine Krankheitsspezifische Anpassung statt (vgl. die Ausführungen weiter oben).

<u>Vorschläge zur Verbesserung/Anpassung der Methodik:</u>
Es soll eine „nachvollziehbare“ und auf die jeweilige Erkrankung abgestimmte Einschätzung des Ausmaßes des Zusatznutzens durch das IQWiG und den G-BA erfolgen. Nur

indikationsspezifische Schwellenwerte können sicherstellen, dass unterschiedliche Ausgangswerte und absolute Risiken Berücksichtigung finden. Auch sollte der Zusatznutzen nicht primär aus statistischen Größen abgeleitet werden.

Deshalb wird eine Berücksichtigung folgender Punkte als sinnvoll angesehen:

- Die Art der Endpunkte, die bei der jeweiligen Erkrankung für den Patienten wichtig sind, sollte berücksichtigt werden. So muss beispielsweise eine eindeutige Zuordnung der Endpunkte in die Zielgrößenkategorien möglich sein. Ebenso ist die Erarbeitung einer nachvollziehbaren Erläuterung der Begriffe „schwerwiegend/schwer“ und „nicht schwerwiegend/ nicht schwer“ notwendig. Besonders gilt dies für Symptome, da für diese keine eindeutige Zuordnung möglich ist, im Gegensatz zu den Nebenwirkungen.
- Auch sollte eine Berücksichtigung der Krankheitsschwere und des natürlichen Verlaufs der Erkrankung stattfinden, sodass gewisse Eigenschaften einer Erkrankung in die Bewertung des Zusatznutzens mit einfließen und somit eine unterschiedliche Gewichtung stattfinden kann. Beispiele sind:
    - Akut vs. Chronisch
    - Letal vs. Nicht-letal
    - Akut vs. Langzeit Effekte/ Komplikationen der Krankheiten
- Die Art der Behandlung/Intervention muss Berücksichtigung in der Bewertung finden, beispielsweise durch eine Unterscheidung zwischen einer Ursachenbehandlung, wo eine Heilung in Aussicht steht oder einer Symptombehandlung, bzw. ob eine kurative oder palliative Behandlung vorliegt
- Zusätzliche Kategorien wie „Heilung“ sollten hinzugefügt werden und nicht wie von IQWiG vorgeschlagen unter die Kategorien Mortalität und Morbidität subsumiert werden.

Diese Art der Berücksichtigung der zugrundeliegenden Erkrankungen kann nicht über ein starres Verfahren geleistet werden; vielmehr muss ein Verfahren institutionalisiert werden, das eine kontinuierliche Anpassung der Methode zur Bewertung des Ausmaßes des Zusatznutzens erlaubt. Es würde sich also um ein dynamisches Bewertungsverfahren handeln, das indikationsspezifisch den Zusatznutzen bewertet, beispielsweise durch ein individuell definiertes Gremium von medizinischen Fachexperten, betroffenen Patienten, Biostatistikern. Die IQWiG-Methoden können einen Rahmen für ein solches Verfahren bilden, indem sie beispielsweise eine Vorkategorisierung der Erkrankungen vornehmen.

Referenzen

- Gemeinsamer Bundesausschuss (2012a), Protokoll „Expertengespräch 22. März 2012: Die frühe Nutzenbewertung von Arzneimittel Umsetzung – Erfahrungen – Folgerungen, www.g-ba.de
- Gemeinsamer Bundesausschusss (2012b), Zusammenfassende Dokumentation über die Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V, Telaprevir vom 29. März 2012, Stand: 21. November 2012, www.g-ba.de
- Heeg B, van Agthoven M, Liwing J et al. Optimal treatment sequencing in multiple myeloma: An exploratory modeling approach. Blood 2010; 116:Abstract 3046.

- IQWiG (2012a), Telaprevir – Nutzenbewertung gemäß § 35a SGB V, Bericht Nr. 115, www.g-ba.de
- IQWiG (2012b), Rilpivirin – Nutzenbewertung gemäß § 35a SGB V, Bericht Nr. 127, www.g-ba.de
- Mühlbacher, A. et al. (2013), Patient preferences for HIV/AIDS therapy - a discrete choice experiment, *Health Economics Review* (forthcoming).

A.1.18 – Lundbeck GmbH

Stellungnahme der Lundbeck GmbH zur Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1 des IQWiG (Entwurf vom 18. April 2013)

Zusammenfassung

- Im vorliegenden Entwurf des IQWiG werden nur ausgewählte Teilbereiche der überarbeiteten oder neuformulierten Methoden veröffentlicht. Es ist jedoch erforderlich, dass auch der Gesamtentwurf des neuen Methodenpapiers zur Diskussion gestellt wird.
- Die vorgeschlagenen Methoden sollten konkreter und ausführlicher erläutert werden, um die Nachvollziehbarkeit zu gewährleisten.
- Es ist erforderlich ein Regelwerk vorzustellen, in dem auch Ausnahmen definiert sind.
- Die Bewertung der klinischen Relevanz von Effekten muss auch mit medizinisch-klinischer Expertise im Indikationsgebiet erfolgen. Eine ausschließlich biometrische Bewertung, die außerdem keinerlei Bezug zu unterschiedlichen Indikationsgebieten hat, ist zur Bewertung des Ausmaßes der klinischen Relevanz nicht geeignet.
- Werteentscheidungen sollen nicht durch das IQWiG getroffen werden. Entsprechende Verfahren können z.B. vom G-BA unter Einbezug unterschiedlicher gesellschaftlicher Gruppen (z.B. Patienten, medizinische Experten) in einem transparenten Verfahren durchgeführt werden.

Allgemeine Anmerkungen

Der vorliegende Methodenentwurf des IQWiG umfasst lediglich wenige Abschnitte bzw. Teilbereiche der zugrunde liegenden „Allgemeinen Methoden 4.0“. Diese Abschnitte wurden vom IQWiG überarbeitet oder durch weitere ergänzt, so dass eine Version „Allgemeine Methoden 4.1“ entstehen wird.

Diese Vorgehensweise ist für Stellungnehmende mit deutlichen Limitierungen verbunden, da nur Teilausschnitte kommentiert werden können ohne jemals die Gesamtschau der zu überarbeitenden Methodik zu diskutieren. Daher ist wichtig, dass auch der Gesamtentwurf des neuen Methodenpapiers „Allgemeine Methoden 4.1“ zur Diskussion gestellt und Stellungnahmemöglichkeit eingeräumt wird.

Des Weiteren ist im Entwurf für den ersten Teilschnitt von Aktualisierungen der Allgemeinen Methoden die Nachvollziehbarkeit nicht vollständig. Es ist notwendig, ein Regelwerk zum methodischen Vorgehen zu erstellen. Exemplarisch werden die Unzulänglichkeiten im Regelwerk in einer Kommentierung von Herrn Prof. Brieden aufgeführt (s. Anlage).

zu Abschnitt „*3.1.4 Endpunktbezogene Bewertung*“

Das IQWiG beabsichtigt, die Aussagesicherheit von Ergebnissen mit den Begriffen „Beleg“, „Hinweis“ oder „Anhaltspunkt“ zu operationalisieren. Hierbei sollen drei Kriterien von Bedeutung sein (s.a. Tabelle 2, Seite 11 des vorliegenden Methodenentwurfs):

- „Anzahl der Studien“
- „qualitative Ergebnissicherheit“
- „Effekt(e)“

Kommentierung zum Begriff „Effekt(e)“

Bei der Bewertung in diesem Kriterium soll zunächst geprüft werden, ob eine Meta-Analyse durchgeführt werden kann (s. auch Kommentierung unter „7.3.8 Meta-Analysen“). Kann nach den Kriterien des IQWiG keine Meta-Analyse durchgeführt werden, sollen die Effekte der jeweiligen Einzelstudien bewertet werden. Das IQWIG beabsichtigt hierbei der „Gleichgerichtetheit“ der Effekte eine besondere Bedeutung beizumessen; entweder, in dem beim Vorliegen von 2 oder 3 Studien die jeweiligen Einzelstudien bewertet werden oder in dem bei 4 und mehr Studien sogenannte „Prädiktionsintervalle“ berechnet werden.

Unklar ist, welcher wissenschaftlich anerkannte Standard diesem Kriterium der „Gleichgerichtetheit“ zu Grunde liegt. Hier sind weitere Ausführungen und Erläuterungen dringend erforderlich.

Unklar ist ebenfalls, wann die „Gleichgerichtetheit“ bewertet werden soll. Im Entwurf der Methoden ist ausgeführt, dass die „Gleichgerichtetheit“ dann bewertet werden soll, wenn wegen zu großer Heterogenität kein gepoolter gemeinsamer Effektschätzer gebildet werden kann – wenn also die Ergebnisse der Einzelstudien nicht in Form einer Meta-Analyse dargestellt werden können.
Auf derselben Seite im Methodenentwurf ist jedoch formuliert, dass das „Prädiktionsintervall“ im Rahmen einer Meta-Analyse mit zufälligen Effekten berechnet werden soll.
Wir bitten das IQWiG daher, diesen Widerspruch aufzulösen.

Kommentierung zum Begriff „qualitative Ergebnissicherheit“

Zur Ergebnissicherheit eines etwaigen Zusatznutzens werden vom IQWiG die Begriffe „qualitative“ und „quantitative“ Ergebnissicherheit eingeführt. Eine Quantifizierung der Ergebnissicherheit ist jedoch lediglich für die qualitative Ergebnissicherheit ansatzweise beschrieben. Eine entsprechende Quantifizierung der quantitativen Ergebnissicherheit wird nicht gegeben. Das IQWiG wird gebeten,

objektivierbare Kriterien zur Quantifizierung der quantitativen Ergebnissicherheit vorzuschlagen.

Kommentierung zum Begriff „Anzahl der Studien“

Im Text ist ausgeführt, dass auch beim Vorhandensein von nur einer Studie ein „Beleg“ abgeleitet werden kann. In der dargestellten Tabelle 2 des vorliegenden Methodenentwurfs (Seite 11) ist diese Option nicht aufgeführt. Die Tabelle 2 drückt aus, dass mindestens 2 Studien vorliegen müssen, um die Bewertung „Beleg“ abzuleiten.

Bei der Bewertung der Aussagesicherheit wird seitens des IQWiG stark auf die Anzahl der klinischen Studien fokussiert. Bei diesem methodischen Ansatz wird nicht die Anzahl der exponierten Patienten berücksichtigt. Diese Bezugsgröße sollte vom IQWIG in der Bewertung der Aussagesicherheit berücksichtigt werden.

Die tabellarische Darstellung wird prinzipiell begrüßt. Wichtig ist jedoch, dass in der Tabelle alle Optionen dargestellt sind. Daher sollte die Tabelle nur dann dargestellt werden, wenn alle Optionen aufgeführt sind.

zu Abschnitt „*7.3.8 Meta-Analysen*“

Im Abschnitt zu Meta-Analysen ist ausgeführt, dass die Darstellung eines gemeinsamen Effektschätzers von der Heterogenität der klinischen Studien abhängig ist. Heterogenität soll über den p-Wert getestet werden, die Quantifizierung mittels des I^2-Maßes erfolgen. In Abhängigkeit dieser Kriterien können Daten einzelner Studien meta-analytisch zusammengefasst werden.

Unklar und teilweise im Widerspruch zu den Ausführungen des IQWiG in Abschnitt „3.1.4 Endpunktbezogene Bewertung“ ist jedoch, wann die sog. „Gleichgerichtetheit“ bewertet werden soll. Im erwähnten Abschnitt ist ausgeführt,

dass „Gleichgerichtetheit“ dann ermittelt werden soll, wenn wegen zu großer Heterogenität kein gepoolter gemeinsamer Effektschätzer gebildet werden kann – wenn also die Ergebnisse der Einzelstudien nicht in Form einer Meta-Analyse dargestellt werden können.
Wie bereits in der Kommentierung zu „3.1.4 Endpunktbezogene Bewertung“ ausgeführt, bitten wir das IQWiG zu erläutern, welcher wissenschaftlich anerkannte Standard dem Kriterium der „Gleichgerichtetheit“ zu Grunde liegt. Außerdem ist es notwendig den bestehenden Widerspruch aufzulösen, „Gleichgerichtetheit“ nur dann darzustellen, wenn keine Meta-Analyse durchgeführt werden kann, jedoch andererseits „Prädiktionsintervalle“ in meta-analytischen Berechnungen zu ermitteln.

zu Abschnitt „*3.1.5 Zusammenfassende Bewertung*“
Das IQWiG führt aus, dass eine quantitative Gewichtung unter Verwendung von Summenscores oder Indizes prospektiv zum Zeitpunkt der Auswahl der zu untersuchenden Endpunkte erfolgen soll. Das IQWiG führt im Weiteren aus, dass geeignete Verfahren zur Gewichtung u.a. der Analytic Hierarchy Process (AHP) oder die Conjoint-Analyse (CA) wären.

Bezüglich der Anwendung der Conjoint-Analyse möchten wir auf die Anmerkungen von Prof. Brieden verweisen; die Kommentierung von Prof. Brieden stellen wir im Anhang zur Verfügung.

Eine Gewichtung von verschiedenen Endpunkten kann bei der Bewertung des Nutzens hilfreich sein. Wichtig ist jedoch, dass bereits mit dieser Gewichtung eine Werteentscheidung verbunden ist. Werteentscheidungen zählen jedoch nicht zum Aufgabenbereich des IQWiG. Daher müssen solche Gewichtungen in einem transparenten Verfahren erfolgen, dass sicherstellt, das durch Beteiligungsrechte alle Perspektiven abgebildet werden können: u.a. die der Erkrankten, die der

medizinischen Experten, die der Gesellschaft. Eine solche Werteentscheidung sollte z.B. vom G-BA durchgeführt werden, nicht jedoch vom IQWiG.

zu Abschnitt „*3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V*“

Das IQWiG beabsichtigt bei der Bewertung nach § 35a SGB V den Nutzen (Zusatznutzen) zunächst für die einzelnen Endpunkte zu bewerten und dann anschließend eine Gesamtbewertung unter Einbezug aller Endpunkte vorzunehmen.

Bewertung der einzelnen Endpunkte:

Im ersten Schritt sollen qualitative Aussagen getroffen werden und die Sicherheit von Ergebnissen mit den Begriffen „Beleg“, „Hinweis“ oder „Anhaltspunkt“ bewertet werden. In der vorliegenden Stellungnahme wurde das vom IQWiG vorgeschlagene Vorgehen konstruktiv diskutiert (Punkte „3.1.4 Endpunktbezogene Bewertung“ und „7.3.8 Meta-Analysen“ dieser Stellungnahme). Daher verweisen wir auf die bereits formulierten Ausführungen. Prinzipiell ist festzustellen, dass die vom IQWiG skizzierte 4-fach Abstufung zur Belegbarkeit des (Zusatz-) Nutzens nicht mit der Verfahrensordnung des G-BA zur Bewertung des Nutzens von Arzneimitteln mit neuen Wirkstoffen nach §35a SGB V korreliert. Wir bitten das IQWiG daher, entsprechend der Verfahrensordnung des G-BA zu verfahren und die 6-fach Abstufung zur Belegbarkeit des (Zusatz-) Nutzens zu übernehmen.

Im zweiten Schritt sollen die dargestellten Effekte bzw. Effektstärken quantifiziert werden. Das IQWiG schlägt hierbei Schwellenwerte vor, mit denen in Abhängigkeit der Zielgrößenkategorie das Ausmaß der Nutzens ermittelt werden soll (s. Tabelle NT1, Seite 18 des vorliegenden Methodenentwurfs).
Diese Vorgehensweise ist medizinisch-wissenschaftlich nicht nachvollziehbar, da z.B. unabhängig von der Erkrankung, dem Schweregrad oder der Versorgungssituation die Quantifizierung des Zusatznutzens mit identischen biometrischen

Schwellenwerten erfolgen soll. Auch nicht nachvollziehbar ist die im Anhang (Seite 26ff. des vorliegenden Methodenentwurfs) erwähnte Rasterung von 1/6.

Die Bewertung des Zusatznutzens oder von Effekten muss in Abhängigkeit der jeweiligen Indikation erfolgen. Daher müssen Kriterien u.a. in Abhängigkeit der jeweiligen Erkrankung, des Schweregrads oder der Versorgungssituation definiert werden. Hierzu muss eine Bewertung unter Einbezug medizinischer Expertise im Indikationsgebiet erfolgen. In Ergänzung können z.B. bei Endpunkten unter Verwendung stetiger Skalen auch biometrische Schwellenwerte abgeleitet werden, die jedoch in direktem Bezug zum Indikationsgebiet stehen müssen.

Unklar ist außerdem, weshalb das IQWiG bei der Bewertung unter Einbezug von Responderkriterien von der bisherigen Vorgehensweise abweicht. Bisher bewertet das IQWiG i.d.R. statistisch signifikante Unterschiede unter Verwendung von akzeptierten Responderkriterien als relevant, da bereits in der Responderdefinition eine MID enthalten ist. Nun beabsichtigt das IQWiG, signifikante Unterschiede auf Basis von Responderkriterien durch die bereits kritisierten biometrischen Schwellenwerte erneut zu bewerten.

Bestehende und klinisch akzeptierte Relevanzdefinitionen wie z.B. Responderdefinitionen sollten auch weiterhin uneingeschränkt durch das IQWiG akzeptiert werden. Sollen biometrische Schwellenwerte ermittelt werden, muss dies indikationsbezogen unter Einbezug medizinischer Expertise erfolgen.

<u>Gesamtschau:</u>

Das IQWiG führt aus, dass die Ergebnisse aller Endpunkte in einer Gesamtschau dargestellt werden sollen und eine Gesamtaussage zum Zusatznutzen getroffen werden soll.

Sollte beabsichtigt sein, Einzeleffekte in der Gesamtschau zu gewichten, trifft das IQWiG Werteentscheidungen. Wie bereits in den Ausführungen zum Punkt „3.1.5

Zusammenfassende Bewertung“ erläutert, gehören Werteentscheidungen nicht zum Aufgabenbereich des IQWiG. Diese müssen in einem offenen und transparenten Dialog unter Einbezug unterschiedlichen gesellschaftlicher Gruppen (z.B. Patienten, medizinische Experten) getroffen werden. Entsprechende Verfahren können vom G-BA initiiert werden.

zu Abschnitt „*Anhang: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens*“

Das IQWiG versucht, eine Auswahl der vom G-BA vorgegebenen Ausmaßkategorien „erheblich“, „beträchtlich“ oder „gering“ zu operationalisieren und mit Inhalten zu besetzen. Diese Ausformulierung sollte jedoch vom G-BA vorgenommen werden und in Form eines transparenten Verfahrens unter Einbezug medizinischer Expertise indikationsbezogen erfolgen. Es ist fest zustellen, dass der G-BA in jüngster Vergangenheit nicht auf die vom IQWiG vorgeschlagene Operationalisierung eingegangen ist, sondern andere Begründungen für die Einstufung der Ausmaßkategorie gegeben hat.

Neben den bereits in der Stellungnahme vorgetragenen Kritikpunkten bitten wir das IQWiG zu erläutern, weshalb relative Effekte grundsätzlich absoluten Effekten vorgezogen werden sollen.

Bei der Festlegung des Ausmaßes eines Zusatznutzens bedient sich das IQWiG einer so genannten „Ausmaßmatrix“, die als eine Rasterung von 1/6-Schritten für die tatsächlichen Effekte definiert wird. Dieses Vorgehen wird als eine „pragmatische Lösung“ (Seite 33) bezeichnet. Wir bitten das IQWiG um eine methodische Begründung für die Rasterung.

Anlage

Kommentierung von Herrn Prof. A. Brieden

Anmerkungen zum Entwurf des IQWIG vom 18.04.2013 zur Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1

Prof. Dr. Andreas Brieden

In einer Vielzahl von unterschiedlichsten Anwendungsfeldern stellen sich problemorientierte Fragen, die nur mit Hilfe quantitativer Methoden fundiert beantwortet werden können. Dabei sind nach Auffassung des Unterzeichnenden diese Methoden grundsätzlich entscheidungsunterstützend einzusetzen, d.h. jedwedes Ergebnis muss stets kontext-abhängig im Rahmen der jeweiligen Anwendung kritisch analysiert werden, um hiernach die richtigen Schlüsse ziehen zu können.

Das vollständige Potential quantitativer Methoden kann folgerichtig nur dann ausgeschöpft werden, wenn sowohl die Methodik als auch die anwendungsspezifische Interpretation auf allerhöchstem Niveau ausgeführt werden, jedoch nicht jeweils isoliert betrachtet, sondern in einem geeigneten Zusammenspiel. Dabei ist darauf zu achten, dass die Methodik an sich Ihre Objektivität beziehungsweise Nachvollziehbarkeit behält. Hierfür ist es notwendig, ein Regelwerk aufzustellen, in welchen Fällen, wie methodisch vorgegangen werden soll. Im Idealfall sind jedoch nicht nur die Regeln, sondern insbesondere auch die Ausnahmen von diesen Regeln zu begründen.

Die Aufstellung eines solchen, fundiert begründeten Regelwerks ist umso schwieriger, desto kontroverser Methoden in der jeweiligen wissenschaftlichen Community diskutiert werden beziehungsweise desto mehr Dynamik in der jeweiligen Forschungsrichtung zu beobachten ist. Eine solche, außerordentlich hohe Dynamik, verbunden mit kontroversen Diskussionen ist im Bereich der evidenzbasierten Medizin zweifelsohne festzustellen. Vor diesem Hintergrund ist der Ansatz des IQWIGs umso begrüßenswerter, ein Regelwerk für den Einsatz quantitativer Methoden aufzustellen. Der eigene Anspruch findet sich beispielsweise auf Seite 10 im Hinweis nach *„regelhaften Anforderungen“* beziehungsweise auf eine

„regelhafte Operationalisierung“ (für die Beleglage bei Effekten). Auch bedarf der *„Begriff der „Heilung“ (…) grundsätzlich einer Operationalisierung, die sich regelhaft auf Kriterien stützen wird“*, S. 29.

Die Aufstellung eines normativen Regelwerks sollte jedoch nicht nur das Ziel verfolgen, ein verwendetes, eigenes Vorgehen im Nachhinein zu begründen. Vielmehr muss auch anderen die Möglichkeit gegeben werden, quantitative Verfahren im Rahmen des vorgeschlagenen Regelwerks mit Aussicht auf Akzeptanz anzuwenden. Und genau an dieser Stelle reicht der vorliegende Entwurf nicht aus. Bei vielen zentralen Punkten wird ein Vorgehen postuliert, dass im Regelfall gelten soll:

*„Die Gewichte der Studien kommen hierbei **in der Regel** aus einer Meta-Analyse mit zufälligen Effekte …“ (Seite 9)*

***„In der Regel** wird an die Aussage eines Belegs die Anforderung zu stellen sein, dass eine Meta-Analyse von Studien mit hoher qualitativer Ergebnissicherheit einen entsprechenden statistisch signifikanten Effekt zeigt.“ (Seite 11)*

*„Eine Meta-Analyse von Studien mit mäßiger qualitativer Ergebnissicherheit oder eine einzelne Studie mit hoher qualitativer Ergebnissicherheit kann trotz statistisch signifikantem Effekt demnach **in der Regel** nur Hinweis liefern.“ (Seite 11)*

*„Eine Meta-Analyse von Studien mit geringer qualitativer Ergebnissicherheit oder eine einzelne Studie mit mäßiger qualitativer Ergebnissicherheit liefert bei statistisch signifikantem Effekt **in der Regel** nur einen Anhaltspunkt.“(Seite 11)*

***„In der Regel** wird von einer statistischen Zusammenfassung abgesehen, falls der Heterogenitätstest einen p-Wert unter 0,2 liefert.“ (Seite 22)*

*„Das Institut interpretiert im Rahmen von Meta-Analysen die Ergebnisse eines Heterogenitäts- oder Interaktionstests bezüglich wichtiger Subgruppen **in der Regel** wie folgt.“ (Seite 24)*

*„Bei Meta-Analysen diagnostischer Studien ist in der Praxis in den meisten Fällen mit Heterogenität zu rechnen, daher empfiehlt sich hier **in der Regel** die Verwendung von Modellen mit zufälligen Effekten.“ (Seite 25)*

*„Das Institut verwendet standardmäßig die übliche Form von Meta-Analysen und greift **in der Regel** nicht auf Methoden für kumulative Meta-Analysen zurück.“ (Seite 26)*

Exakt zu beschreiben, wie in der Regel vorgegangen werden soll ist zweifelsohne notwendig für die Akzeptanz der Anwendung der quantitativen Verfahren. Wann kann, soll oder muss jedoch von diesen Regeln abgewichen werden und wie ist dann zu verfahren?

So bleiben an verschiedenen Stellen (siehe obige Zitate) und zu den verschiedenen Methoden zentrale Fragen unbeantwortet.

Bei der Durchführung einer Meta-Analyse im Rahmen der endpunktbezogenen Bewertung (s. 3.1.4) ist die Sensitivitätsanalyse ein geeignetes Instrumentarium, um die Ergebnissicherheit zu untersuchen. Allerdings, nach welchen Regeln wird diese durchgeführt? Werden etwa im Fall der Meta-Analyse unter Annahme zufälliger Effekte bayesianische und nicht-bayesianischen Ansätze verglichen und nach welchen Regel werden die Prior-Verteilungen variiert?

Auch die zusammenfassende Bewertung (s. 3.1.5) mit Hilfe multikriterieller Entscheidungsverfahren wie etwa dem Analytic Hierarchy Process (AHP) oder der Conjoint-Analyse (CA) ist ein begrüßenswerter Ansatz. Für seine Akzeptanz bzw. eigenständige Durchführbarkeit ist aber unter anderem zu beantworten, welche Fallzahlen beispielsweise bei der Durchführung des AHP verlangt werden? Und nach welchem Regelwerk soll bei Anwendung einer Conjoint-Analyse gewährleistet werden, dass die Auswahl der Stimuli derart erfolgt, dass die aus der Analyse resultierende Gewichtung hierdurch nicht manipuliert werden kann? Und ergeben sich die Gewichte aus einer aggregierten Regression oder der Mittelung einzelner Regressionen?

Zusammenfassend ist als Fazit festzuhalten, dass der vorgelegte Entwurf der Bedeutung der quantitativen Methoden gebührend Rechnung trägt, was nach Auffassung des Unterzeichnenden mehr als begrüßenswert ist. Die extrem schwierige Aufgabe, in einem komplexen Umfeld ein normatives Regelwerk aufzustellen, ist jedoch nur unvollständig gelöst; hierfür sind wesentlich detailliertere Ausführungen erforderlich.

Abschließend sei an die einleitende Bemerkung erinnert, dass quantitative Methoden immer nur entscheidungsunterstützend sein können und immer anwendungsbezogenes Expertenwissens notwendig ist, um die richtigen Schlüsse zu ziehen. Von daher sollten auch Parameter und Schwellenwerte zwar konzeptionell in der Methodik vorgesehen werden, **in der Regel** jedoch fall- beziehungsweise anwendungsspezifisch von Fachexperten auf dem Gebiet der Biometrie/Mathematik und medizinischen Experten festgelegt werden.

Prof. Dr. Andreas Brieden München, 13. Mai 2013

A.1.19 – MSD Sharp & Dohme GmbH

MSD SHARP & DOHME GMBH
Lindenplatz 1
85540 Haar
[redacted]
www.msd.de

MSD

MSD SHARP & DOHME GMBH · Postfach 12 02 · 85530 Haar
Institut für Qualität und Wirtschaftlichkeit
im Gesundheitswesen
Herrn Prof. Dr. med. Jürgen Windeler
Im Mediapark 8
50670 Köln

22. Mai 2013

Stellungnahme der MSD SHARP & DOHME GMBH
zum Entwurf der Version 4.1 der Allgemeinen Methoden vom 18.04.2013

Sehr geehrter Herr Professor Windeler,

wir danken für die Möglichkeit der Kommentierung des Entwurfs.

Aus Gründen der Übersichtlichkeit und der Abbildung eines Einzelstudienbelegs empfehlen wir, auf Seite 11 des Entwurfes die nachfolgende Tabelle

Tabelle 2: Anforderungen an die Beleglage für die unterschiedlichen Aussagesicherheiten beim Vorliegen von Studien derselben qualitativen Ergebnissicherheit

	Anforderung		
Aussage	**Anzahl der Studien**	**qualitative Ergebnissicherheit**	**Effekt(e)[a]**
Beleg	≥ 2	hoch	homogen, Meta-Analyse statistisch signifikant
	≥ 2	hoch	heterogen deutlich gleichgerichtet
Hinweis	≥ 2	mäßig	homogen, Meta-Analyse statistisch signifikant
	> 2	mäßig	heterogen deutlich gleichgerichtet
	> 2	hoch	heterogen mäßig gleichgerichtet
	1	hoch	statistisch signifikant
Anhalts-punkt	> 2	gering	homogen, Meta-Analyse statistisch signifikant
	≥ 2	gering	heterogen deutlich gleichgerichtet
	≥ 2	mäßig	heterogen mäßig gleichgerichtet
	1	mäßig	statistisch signifikant

a: Zur Erläuterung des Begriffs: siehe Text.

Geschäftsführer:
Hanspeter Quodt (Vorsitzender)
Dr. Veit Stoll
Dr. Claus Dollinger
Aufsichtsratsvorsitzender:
Dr. Axel Breitstadt

[redacted]

Sitz der Gesellschaft:
Lindenplatz 1
85540 Haar
Tel. +49 89 45 61 - 0
Fax +49 89 460 10 10
www.msd.de

Infocenter:
E-Mail infocenter@msd.de
Tel. 0800 673 673 673
Fax 0800 673 673 329

zu ersetzen durch:

Tabelle 2: Anforderungen an die Beleglage für die unterschiedlichen Aussagesicherheiten beim Vorliegen von Studien derselben qualitativen Ergebnissicherheit

		Anzahl Studien				
		1	≥2			
			Meta-Analyse statistisch signifikant	Heterogen		
				Gleichgerichtete Effekte		
				Deutlich	Mäßig	Nein
Qualitative Ergebnis-sicherheit	Hoch	Beleg/Hinweis	Beleg	Beleg	Hinweis	-
	Mäßig	Anhaltspunkt	Hinweis	Hinweis	Anhaltspunkt	-
	Gering	-	Anhaltspunkt	Anhaltspunkt	-	-
Erläuterungen siehe Text						

Wir begründen dies wie folgt.

In der Tabelle 2 des derzeitigen Entwurfs fehlt die im Text beschriebene Konstellation des Einzelstudienbelegs:

„Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und ihre Ergebnisse besondere Anforderungen zu stellen".

Nicht abschließende Beispiele für solche Konstellationen sind aus unserer Sicht:

- Die im Entwurf genannten "Points to consider" der EMA (Ihre Referenz 157).
- Dramatische Effekte aus einer (1) randomisierten Studie hoher qualitativer Ergebnissicherheit im Direktvergleich mit einer zweckmäßigen Vergleichstherapie.
- Unter "4.1 Anwendungsgebiete" liegen Beschlüsse der EMA zu patientenrelevanten Endpunkten vor, beispielsweise "indiziert zur Senkung der kardiovaskulären Mortalität und Morbidität", wiederum basierend auf einer (1) randomisierten Studie hoher qualitativer Ergebnissicherheit im Direktvergleich mit einer zweckmäßigen Vergleichstherapie.

MSD

Nur ein "Hinweis" auf einen Zusatznutzen für solche Endpunkte würde regulatorische Beschlüsse konterkarieren und dem Erkenntnisgewinn aus Meilensteinstudien nicht gerecht.

Wir bitten daher, die Möglichkeit eines Einzelstudienbelegs in die Darstellung mit aufzunehmen.

Mit freundlichen Grüßen
MSD SHARP & DOHME GMBH

Dr. Dr. Karl J. Krobot
Direktor Outcomes Research / HTA

Dr. Monika Scheuringer
Senior Manager Biostatistics und
HTA Management

A.1.20 – Novartis Pharma GmbH

NOVARTIS

Stellungnahme der Novartis Pharma GmbH zum Entwurf des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) in der Fassung vom 18.04.2013 zur

„Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“

Verfasst von Sven Klebs, Steffen Jugl, Katja Neidhardt, Volker Claus, Christiane Hohmann, Timo Wasmuth, Renate Handrock, Christian Sieder – alle Novartis Pharma GmbH, Nürnberg

Die Novartis Pharma GmbH begrüßt die Möglichkeit zum oben genannten Entwurf zur Aktualisierung und Ergänzung der Allgemeinen Methoden des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen Stellung zu nehmen. Die folgenden Punkte des aktuellen Entwurfs möchten wir in diesem Rahmen gerne kommentieren, da sie einer eingehenden Erörterung bedürfen:

1 Vorschlag des IQWiG zu Möglichkeiten zur mündlichen Erörterung von IQWiG - Berichten (Abschnitt 2.1.1 und 2.2.3.)

In seinem neuen Entwurf des Methodenpapiers sieht das IQWiG in Abschnitt 2.1.1 die mündliche wissenschaftliche Erörterung bei der Erstellung des Berichtsplanes und des Vorberichtes weiterhin als lediglich „*optional*“ vor. Es liegen keinerlei transparente Kriterien vor, nach welchen das IQWiG entscheidet, ob eine entsprechende Erörterung erfolgt oder nicht. **Im Sinne einer möglichst breiten Einbeziehung von Expertise sind mündliche Erörterungen daher zwingend zu fordern.**

Im aktuell gültigen Methodenpapier 4.0 des IQWiG ist folgender Prozessschritt enthalten, den das IQWiG zukünftig nun nur noch optional durchführen möchte: *„Zudem wird als weiterer Schritt der Qualitätssicherung der Vorbericht einem oder mehreren externen Reviewern (siehe Abschnitt 2.2.3) mit ausgewiesener methodischer und / oder fachlicher Kompetenz vorgelegt.*“ **Es ist nicht nachzuvollziehen, warum das IQWiG nicht weiterhin eine obligate externe Qualitätssicherung vorsieht. Eine solche externe Qualitätssicherung ist weiterhin obligat erforderlich.**

Dies bildet sich auch entsprechend im Abschnitt 2.2.3 des Methodenpapiers ab. Bisher galt: *„Alle Produkte einschließlich der jeweiligen Zwischenprodukte unterliegen einem umfangreichen mehrstufigen internen Qualitätssicherungsverfahren. Darüber hinaus wird im Verlauf der Erstellung von Berichten und z. T. auch von Gesundheitsinformationen ein externes Reviewverfahren als weiterer Schritt der Qualitätssicherung durchgeführt.“* Im aktuellen Methodenentwurf des IQWiG ist dies, wie oben dargestellt, nur noch als optionaler Schritt vorgesehen. Die Durchführung eines externen Reviewverfahrens muss jedoch weiterhin obligat erfolgen.

Ferner ist zu kritisieren, dass das IQWiG die externen Reviews weiterhin nicht veröffentlichen möchte. Im Sinne der Transparenz ist dies nicht nachzuvollziehen. Die Bewertung durch externe Experten ist für die Akzeptanz und wissenschaftliche Diskussion ein wichtiger Bestandteil. **Analog zu den Stellungnahmen ist daher auch der externe Review zu publizieren. Mindestens ist zu fordern, dass Art und Inhalt der an die Experten gerichteten Fragen und die darauf erfolgten Antworten dargestellt werden.**

2 Vorschlag zur Operationalisierung von Ergebnissicherheit und –ausmaß zur endpunktbezogenen Bewertung (Abschnitt 3.1.4)

Im Abschnitt 3.1.4 seines Methodenentwurfs beschreibt das IQWiG die Anforderungen an die erforderliche Studienlage, um aus der verfügbaren Evidenz auf Endpunktebene einen ‚Beleg‘, einen ‚Hinweis‘ oder einen ‚Anhaltspunkt‘ konstatieren zu können. Hierzu nennt das IQWiG folgende drei Prüfkategorien: Anzahl der Studien, qualitative Ergebnissicherheit der Studien und Effekt auf Endpunktebene. Das IQWiG definiert hierfür weitere Bewertungskriterien für die Einordnung der Beleglage auf Basis der drei Prüfkategorien (vgl. Tabelle 2 im Entwurf des Methodenpapiers). Für einen ‚Beleg‘ müssen im Regelfall demnach mindestens zwei Studien von hoher qualitativer Ergebnissicherheit mit statistisch signifikanten Endpunkteffekten vorliegen, die homogen sind oder - sofern heterogen -

deutlich gleichgerichtet sein müssen. Deutlich gleichgerichtet definiert das IQWiG dahingehend, dass ein Studienpool von mindestens zwei Studien mit jeweils statistisch signifikantem Effekt vorliegen muss, deren Ergebnisse nicht durch weitere vergleichbare ergebnissichere Studien infrage gestellt werden. In Konsequenz bedeutet dies, dass - liegen nur zwei für die Bewertung relevante Studien mit heterogener Ergebnislage vor und ist in einer der beiden Studien der endpunktbezogene Effekt nicht statistisch signifikant - eine solche Ergebnislage regelhaft vom IQWiG nicht als ‚Beleg' gewertet wird (beispielsweise der Fall, wenn neben einer großen multizentrischen Studie hoher Fallzahl mit signifikantem Effekt eine weitere kleinere Studie mit niedriger Fallzahl und statistisch nicht signifikantem Ergebnis vorliegt). Im neuen Methodenentwurf sieht das IQWiG ebenfalls eine Abwertung der Ergebnissicherheit für den Fall vor, dass nur eine einzelne Studie mit hoher qualitativer Ergebnissicherheit vorliegt und führt aus, dass eine einzelne Studie mit hoher qualitativer Ergebnissicherheit trotz statistisch signifikantem Effekt demnach in der Regel nur einen ‚Hinweis' liefern kann. *„Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und ihre Ergebnisse besondere Anforderungen zu stellen.*" Dabei beruft sich das IQWiG auf ein Empfehlungspapier der Europäischen Zulassungsbehörde EMA, wonach bei Zulassungsverfahren, die auf nur einer pivotalen Studie beruhen, an diese Studie besondere Anforderungen seitens der EMA gestellt werden [1]. Bei seiner Auslegung lässt das IQWiG dabei außer Acht, dass seitens der EMA durch die Erteilung einer Zulassung für ein neues Arzneimittel auf Basis einer einzigen Studie, die speziellen EMA-Anforderungen an die Qualität dieser einen Studie und deren Evidenzlage erfüllt sind.

Aus Sicht der Novartis Pharma GmbH liegt auf Basis bereits nur einer pivotalen Studie eine ausreichende qualitative Ergebnissicherheit für den „Beleg" eines (Zusatz-) Nutzens auf Endpunktebene grundsätzlich vor, sofern die Zulassung des zu bewertenden Arzneimittels durch die EMA auf Basis dieser einen pivotalen Studie erfolgte. Mit der Zulassung sind die vom IQWiG geforderten hohen qualitativen Anforderungen der EMA an diese Einzelstudie gegeben.

Das IQWiG unterscheidet zu Recht zwischen dem Ausmaß und der Ergebnissicherheit des Zusatznutzens und verwendet für diese beiden unterschiedlichen Dimensionen auch unterschiedliche Klassifizierungen (‚erheblich', ‚beträchtlich', ‚gering', ‚nicht quantifizierbar' für das Ausmaß bzw. ‚Beleg', ‚Hinweis', ‚Anhaltspunkt' für die Ergebnissicherheit). Umso mehr verwundert das Konzept, für die Klassifizierung des Ausmaßes eines Effekts nicht den Punktschätzer, sondern die obere Grenze des Konfidenzintervalls zu verwenden. Der Punktschätzer wird in der Literatur zu Recht als der ‚best guess' bezeichnet, er ist die ‚wahrscheinlichste', d.h. die plausibelste Schätzung für die Größe des wahren Effekts. Das Konfidenzintervall dagegen beschreibt die (Un-)Sicherheit dieser Abschätzung, gehört also logisch zur Ergebnissicherheit. Das IQWiG vermischt die Kategorien ‚Ergebnissicherheit' und ‚Ausmaß', welche es vorher logisch getrennt hatte und bringt die Ergebnissicherheit dann gleich doppelt in die Bewertung ein: als Forderung nach statistischer Signifikanz, um zumindest einen ‚Anhaltspunkt' bei der Ergebnissicherheit anzuerkennen und dann noch einmal als obere Grenze des Konfidenzintervalls zur Beurteilung der Effektstärke.

An dieser Stelle sei ebenfalls auf die Inflationierung und Neudefinition etablierter Begrifflichkeiten durch das IQWiG hingewiesen: Im üblichen Sprachgebrauch ist ein signifikantes Ergebnis einer konfirmatorischen Studie ein statistischer Beleg. Ein Hinweis oder ein Anhaltspunkt wäre beispielsweise ein nicht-signifikanter Trend, ein Ergebnis für

einen Surrogatendpunkt oder einer unkontrollierten Proof-of-Concept-Studie. Diese Hinweise lässt das IQWiG nicht einmal als ‚Anhaltspunkt' gelten, sondern erklärt sie gleich für nicht ausreichend relevant. Stattdessen wird statistische Signifikanz als Mindestbedingung selbst für einen ‚Anhaltspunkt' gefordert.

Aus dieser Vermischung der Kategorien Ergebnissicherheit und Effektstärke ergeben sich dann weitere Paradoxa: So ist in der Statistik üblicherweise die Effektstärke unabhängig von der Studiengröße (bzw. von der Anzahl der Studien bei Metaanalysen). Diese bestimmt stattdessen nur die Genauigkeit der Abschätzung, also die Ergebnissicherheit. Nach der IQWiG-Definition dagegen hängt die Effektstärke (auch) von der Fallzahl ab. Demzufolge würde man für die Effektkategorien ‚beträchtlich' und ‚erheblich' deutlich höhere Patientenzahlen benötigen als zum einfachen Nachweis der Überlegenheit eines Arzneimittels erforderlich wären. Die ethische Rechtfertigung und Begründbarkeit solcher Fallzahlen in klinischen Studien, die ja auch immer mit einem Risiko für die Patienten verbunden sind, bleibt fragwürdig.

Das IQWiG geht bei der Festlegung der Schwellenwerte für die Effektkategorien vom Vorliegen zweier identischer pivotaler Studien aus, die für eine einfache Überlegenheit gepowert wurden. Diese Annahme ist ausgesprochen unrealistisch. Bei Indikationen mit den Endpunkten Mortalität/Morbidität ist den Zulassungsanforderungen der EMA mit einer pivotalen Studie in der Regel genüge getan. Große Endpunktstudien werden zudem fast immer aus ethischen Erwägungen mit Interimsauswertungen durchgeführt und ggf. frühzeitig abgebrochen, sobald der Vorteil der neuen Therapie nachgewiesen ist, also sobald das Konfidenzintervall unterhalb von Null endet. Damit ist nach der IQWiG-Klassifikation niemals ein höherer als ein ‚geringer' Effekt nachweisbar. Und selbst wenn zwei pivotale Studien vorliegen, so richten sich diese im Allgemeinen an den – mit den Behörden für die Studiendurchführung abgestimmten – Anforderungen für die Zulassung aus und erfüllen dann nicht unbedingt gleichzeitig auch die IQWiG-Kriterien bzgl. Endpunkten, Komparatoren und Strukturgleichheit, können also für die Nutzenbewertung oft gar nicht verwendet werden. Hinzu kommt die Problematik der Zerlegung einer Studienpopulation in mehrere Teilpopulationen. Während die Zulassungsbehörden explizit dazu auffordern, bei zwei pivotalen Studien nicht einfach nur die erste zu reproduzieren, sondern dabei ein möglichst breites Spektrum klinischer Designparameter wie beispielsweise ‚disease conditions', Subpopulationen, Vor- und/oder Begleitmedikation, Komparatoren abzudecken [1], werden diese Variationen vom IQWiG und dem Gemeinsamen Bundesausschuss als Anlass genommen, die Gesamtpopulation in Teilpopulationen aufzuteilen, und dann für jede Teilpopulation den vollen statistischen Evidenznachweis zu fordern. Auch die bisher vorliegenden Nutzenbewertungen haben gezeigt, dass die Annahme einer ‚Verdopplung der Fallzahl' zur Ableitung der Schwellenwerte absolut unrealistisch ist. Durch die – von den Zulassungskriterien abweichenden – Anforderungen des IQWiG und durch die praktizierte Zerlegung in die Anwendungsgebiete geht meistens nur eine Teilmenge der Patienten einer Zulassungsstudie in die Nutzenbewertung ein. Die besseren Effektkategorien ‚beträchtlich' und ‚erheblich' sind somit praktisch nur selten erreichbar.

Binäre Zielgrößen

Das Methodenpapier geht davon aus, dass die Effekte eines Arzneimittels bei binären Zielgrößen prinzipiell als relatives Risiko dargestellt werden. Die Ausführungen im Anhang über die Vorzüge relativer Maße im Vergleich zu absoluten (Risikodifferenz) werden

ausdrücklich begrüßt. Allerdings liegen die Ergebnisse klinischer Studien häufig als Odds Ratio vor. Diese sind das Ergebnis von Auswertungen mittels logistischer Regression, welche die Berücksichtigung prognostischer Faktoren erlaubt und daher ein deutlich genaueres und unverfälschteres Ergebnis als eine unadjustierte Berechnung des relativen Risikos liefert. Es bleibt unklar, ob die Odds Ratio als alternatives relatives Effektmaß anerkannt wird und ob dafür dann dieselben Relevanzgrenzen gelten sollen.

Stetige (quantitative) Zielgrößen

Bei stetigen Zielgrößen sieht das Methodenpapier die Dichotomisierung anhand eines etablierten Responderkriteriums als einzige Methode zur Beurteilung der Effektstärke vor. Auch wenn sich dieses Verfahren einiger Beliebtheit erfreut, so erscheint es als alleinig angewandtes logisch widersinnig, ist statistisch ineffektiv und klinisch undifferenziert.

- Logisch ist es widersinnig, zur Quantifizierung einer quantitativen(!) Zielgröße zu dichotomisieren und sie damit ihrer quantitativen Eigenschaft zu berauben, um danach über Schwellenwerte für die Responderraten wieder eine Quantifizierung einzuführen.
- Statistisch ist es ineffektiv, weil die Dichotomisierung zu einem massiven Verlust von statistischer Power führt und die meisten Zulassungsstudien auf die quantitative Auswertung der stetigen Zielgröße gepowert, also für die dichotome Responderanalyse unterpowert sind. Zudem ermöglicht eine Responderanalyse keine Adjustierung bezüglich relevanter prognostischer Faktoren, was zu weiteren Unschärfen und Powerverlusten führt.
- Klinisch ist es undifferenziert, weil es Patienten nur in die zwei Kategorien ‚Responder' und ‚Non-Responder' einteilt und somit z.B. einen Patienten, der das Responsekriterium knapp erfüllt von einem – sehr ähnlichen – Patienten, der dieses Kriterium knapp verfehlt, unterscheidet, letzteren aber mit Patienten, die sich gar nicht verbessern oder gar verschlechtern, gleichsetzt.

Für viele Zielgrößen sind zudem keine Responsekriterien definiert, oder wenn, dann sind diese oft umstritten und/oder bei genauerer Betrachtung widersprüchlich, beziehungsweise inkonsistent.

Aus Sicht der Novartis Pharma GmbH soll sinnvollerweise eine Responderanalyse, im Falle vorliegender validierter bzw. etablierter Responsekriterien, als gleichwertig zur Analyse der (adjustierten) Differenzen mittels ANCOVA angesehen und durchgeführt werden. Das Ausmaß des Zusatznutzens ist dabei anhand aller vorliegenden Analysen klinisch und indikationsspezifisch zu bewerten.

3 Vorschläge des IQWiG zur Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V (Abschnitt 3.3.3)

3.1 Vorschlag über die zugrunde zu legenden Standards

Zu Beginn des Abschnitt 3.3.3. im Methodenentwurf zur Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V bezieht sich das IQWiG auf die Dossiervorlagen, d.h. auf die Vorgaben des Gemeinsamen Bundesausschuss für die erforderlichen Dossiers zur Nutzenbewertung

von Arzneimitteln gemäß § 35a SGB V, die in der Verfahrensordnung des Gemeinsamen Bundesausschuss geregelt sind [2]. Mit Beschluss vom 18. April 2013 hat der Gemeinsame Bundesauschuss diese Dossiervorlagen in wesentlichen Punkten (z. B. in den Vorgaben zur Kostendarstellung, s.u.) geändert [3]. Das IQWiG bezieht sich in seinem aktuellen Methodenentwurf vom 18.04.2013 aber noch auf die Vorgängerversionen der Dossiervorlagen. Durch die zeitliche Überlagerung von Änderungen in den Dossiervorlagen und im Methodenpapier, ist es nicht transparent und sachgerecht, möglich Veränderungen nachzuvollziehen und zu kommentieren. **Aus diesem Grund schlägt die Novartis Pharma GmbH vor, dass das IQWiG den vorliegenden Methodenentwurf zunächst an die zukünftig gültigen Rahmenbedingungen (Dossiervorlagen) anpasst und danach der Öffentlichkeit erneut zur Kommentierung zur Verfügung stellt.**

Im Gegensatz zu § 35a SGB V, in dem es heißt, dass „*die **internationalen** [Hervorhebung durch den Verfasser] Standards der evidenzbasierten Medizin und der Gesundheitsökonomie*“ Grundlage für den Prozess der frühen Nutzenbewertung sind, schreibt das IQWiG im Methodenentwurf nur von „*Standards*“ der evidenzbasierten Medizin und der Gesundheitsökonomie.

Der Gesetzgeber hat hier bewusst den Terminus ‚internationale Standards‘ und nicht ähnliche Termini, wie „*allgemein anerkannten wissenschaftlichen Standards*“ (vgl. § 65 SGB V) oder „*wissenschaftliche Begleitung und die Auswertung*“ (vgl. § 137e SGB V) oder „*allgemein anerkannter wissenschaftlicher Standards*“ (vgl. § 137f SGB V) verwendet.

Die Novartis Pharma GmbH weist darauf hin, dass es sich bei der vom IQWiG angewendeten Methodik nur teilweise um internationale Standards handelt. Vielmehr ist die Methodik eine Eigenentwicklung des IQWiG, die sich zwar aus einem internationalen Methodenpool bedient, sich aber in seiner Zusammen- und Umsetzung gänzlich von den internationalen Standards und deren Anwendung unterscheidet. Daher bedarf in diesem Kontext die Entwurfsfassung der Allgemeinen Methoden 4.1 des IQWiG der weiteren Präzisierung durch die Bezugnahme auf entsprechende internationale gesundheitsökomische Methodik.

3.2 Vorschlag zur Bewertung von Kosten

Im Abschnitt 3.3.3. des Methodenentwurfs schreibt das IQWIG, dass die Kosten für das zu bewertende Arzneimittel und die zweckmäßige Vergleichstherapie gemessen am Apothekenabgabepreis anzugeben sind (vgl. S. 15 des Methodenentwurfs). Demgegenüber steht die geforderte Operationalisierung des Gemeinsamen Bundeausschusses in der aktuellen Fassung der Dossiervorlagen zur Kostendarstellung [3]. Diese widersprechen den Ausführungen des IQWiG.

Aus Sicht der Novartis Pharma GmbH erfüllt die im vorliegenden Methodenentwurf - und im Übrigen auch die in den aktualisierten Dossiervorlagen zur Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V - dargestellte Methodik, wie oben bereits erwähnt, nicht die gesetzlichen Vorgaben aus § 35a Abs.1 SGB V und §7 Abs. 2 AM-NutzenV, wonach für die Bewertung der Therapiekosten die internationalen Standards der Gesundheitsökonomie zu befolgen sind.

In der aktualisierten Dossiervorlage heißt es zum Abschnitt 3.3.3.:

„Generell soll(en) die für die Behandlungsdauer zweckmäßigste(n) und wirtschaftlichste(n) verordnungsfähige(n) Packungsgröße(n) gewählt werden. Sofern Festbeträge vorhanden sind, müssen diese angegeben werden. Sofern keine Festbeträge bestehen, soll das günstigste Arzneimittel gewählt werden. Importarzneimittel sollen nicht berücksichtigt werden. Geben Sie zusätzlich die den Krankenkassen tatsächlich entstehenden Kosten an.“ [4].

Mit der obligaten Angabe des günstigsten Arzneimittels wird die implizite Annahme getroffen, dass an alle Patienten der Zielpopulation dieses spezifische Arzneimittel abgeben wird. Dies entspricht jedoch nicht den tatsächlichen Kosten der Krankenkassen. Die maßgeblichen und rechtsverbindlichen Vorschriften zur Abgabe und daraus resultierenden Kosten für die Krankenkassen werden nicht durch die Vorgaben in der Dossiervorlage geregelt, sondern im aktuellen Rahmenvertrag über die Arzneimittelversorgung nach § 129 Abs. 2 SGB V [5]. In diesem werden u.a. diverse Kriterien angeführt, unter welchen Umständen welches Arzneimittel als das preisgünstigste und damit wirtschaftlichste abzugeben ist.

Die Nichtberücksichtigung dieses Rahmenvertrages ist ein Widerspruch in sich, da im Folgesatz der Dossiervorlage gefordert wird, gerade diese tatsächlichen, den Krankenkassen entstehenden Kosten anzugeben. Weiterhin steht die Beschränkung auf das günstigste Arzneimittel als einzig zulässigen Punktschätzer der expliziten Forderung des IQWiG und des G-BA nach der Angabe von Spannen mit Unter- und Obergrenzen entgegen [4].

Aus Sicht der Novartis Pharma GmbH spiegeln nach Marktanteilen gewichtete Kosten ein den tatsächlichen Kosten der Krankenkassen realistischeres Bild wider, da alle abgegebenen und abgerechneten Arzneimittel den Kriterien des oben genannten Rahmenvertrags folgen müssen.

Auch Internationale gesundheitsökonomische Expertengremien, wie z. B. die ‚Task Force on Good Research Practices – Budget Impact Analysis‘ der International Society for Pharmacoeconomics and Outcomes Research (ISPOR), fordern in ihren Methodenpapieren explizit nach realistischen und gerechtfertigten Annahmen und einer Berücksichtigung des „treatment mix“ bei der Berechnung von Kostenauswirkung auf die Budgets der jeweiligen Kostenträger [6]. Dieser ‚Treatment Mix‘ bezieht sich in diesem Methodenpapier nicht auf Wirkstoffebene, sondern explizit auf Arzneimittel („*drugs*“ anstatt INN oder „*substance*“). Dieses findet sich auch in einer kürzlich veröffentlichten Draft-Version eines geplanten weiteren Methodenpapieres der gleichen Task Force wieder [7]. Auf nationaler Ebene wird das oben beschriebene Vorgehen ebenfalls vorgeschlagen. Beispielhaft seien hier die Publikationen von Graf von Schulenburg et al. 2007 und Braun et al. 2009 genannt [8, 9, 10].

3.3 Vorschlag zur Operationalisierung der Quantifizierung des Ausmaßes des Zusatznutzens

Im Abschnitt 3.3.3. auf Seite 15 ff. sowie im Anhang der Entwurfsfassung der Allgemeinen Methoden 4.1 beschreibt das IQWIG, dass im Rahmen der Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V das Ausmaß des Zusatznutzens gemäß den Kategorien der Arzneimittel-Nutzenbewertungsverordnung beschrieben werden soll (erheblicher, beträchtlicher, geringer, nicht quantifizierbarer Zusatznutzen, kein Zusatznutzen belegt, Nutzen des zu bewertenden Arzneimittels geringer als Nutzen der zweckmäßigen Vergleichstherapie). Hierfür sieht das IQWiG in seiner jetzt vorgelegten Entwurfsfassung der Allgemeinen Methoden 4.1 vom 18.04.2013 eine Vorgehensweise zur Feststellung des Zusatznutzens in drei Schritten vor:

- Schritt 1 (qualitative Aussage): für jeden Endpunkt separat die Prüfung der Wahrscheinlichkeit für das Vorliegen eines Effekts. Je nach Güte der Evidenz wird die Wahrscheinlichkeit als ‚Anhaltspunkt', ‚Hinweis' oder ‚Beleg' eingestuft.
- Schritt 2 (quantitative Aussage): für die Endpunkte, für die im Schritt 1 zumindest ein Anhaltspunkt für das Vorliegen eines Effekts attestiert wurde, wird jeweils separat das Ausmaß der Effektstärke festgestellt. Dabei sind folgende quantitative Aussagen möglich: ‚erheblich', ‚beträchtlich', ‚gering', ‚nicht quantifizierbar'.
- Schritt 3 (Gesamtaussage zum Zusatznutzen): im Rahmen einer Gesamtschau wird anhand aller Endpunkte unter Würdigung der Wahrscheinlichkeit und des Ausmaßes auf Endpunktebene die Gesamtaussage zum Zusatznutzen festgestellt.

Zur Festlegung der quantitativen Aussage (Schritt 2) auf Endpunktebene schlägt das IQWiG in Tabelle NT1 der Entwurfsfassung der Allgemeinen Methoden 4.1 Schwellenwerten für relative Effektmaße dar und führt zur Methodik der Anwendung der Schwellenwerte auf Seite 16 weiter aus: *„Das grundsätzliche Konzept sieht vor, für relative Effektmaße Schwellenwerte für Konfidenzintervalle in Abhängigkeit von anzustrebenden Effekten abzuleiten, die wiederum von der Qualität der Zielgrößen und den Ausmaßkategorien abhängen."*

Aus Sicht der Novartis Pharma GmbH sind die vom IQWiG in Tabelle NT1 zur Festlegung des Ausmaßes des Zusatznutzens vorgeschlagenen Schwellenwerte sowie die dargestellte Operationalisierung zur Bestimmung des Ausmaßes des Zusatznutzens abzulehnen, weil

1.) **Die vorgeschlagenen Schwellenwerte in ihrer Herleitung weder evidenz-basiert noch werturteil-begründet sind und vom IQWiG ohne die Einbindung der Sichtweise von Betroffenen nicht getroffen werden kann.**
2.) **Die Anwendung der Operationalisierung im Rahmen der Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V nicht akzeptiert ist, insofern da sie bis heute in den Beschlüssen des Gemeinsamen Bundesausschusses bei der Entscheidungsfindung zur Quantifizierung des Zusatznutzenausmaßes nicht berücksichtigt wurde.**

Bei den in Tabelle NT1 dargestellten Schwellenwerten für die Bestimmung des Ausmaßes des Zusatznutzens handelt es sich um die vom IQWiG im Zuge des ersten

Nutzenbewertungsverfahrens nach § 35a SGB V von Ticagrelor erstmalig vorgeschlagenen und angewendeten und im Rahmen dieses Verfahrens zur weiteren Diskussion gestellten Schwellenwerte für die Methodik zur Kategorisierung des Zusatznutzens nach AM-NutzenV [11]. Im vorgelegten Methodenentwurf werden dabei vom IQWiG weder die durch verschiedene Stellungnehmenden im Ticagrelor-Verfahren gegenüber dieser Methodik bereits vorgebrachten medizinisch-ethischen Bedenken aufgegriffen oder berücksichtigt, noch würdigt das IQWiG darin den Sachverhalt, dass der Gemeinsame Bundesausschuss bei der Bestimmung des Ausmaßes des Zusatznutzens bereits bei der Bewertung von Ticagrelor sowie in den nachfolgenden Verfahren zur Nutzenbewertung nach § 35a SGB V auf die vom IQWiG vorgeschlagene Methodik bis heute nicht abgestellt hat [12].

Gegen die vom IQWiG vorgeschlagene Methodik und Vorgehensweise zur Festlegung der Schwellenwerte lässt sich insbesondere folgendes einwenden:

1.) **Die Kategorisierung eines Großteils der verschiedenen patienten-relevanten Endpunkte und die Festlegung der unterschiedlichen Schwellenwerte für die Bestimmung des Zusatznutzenausmaßes ist weder evidenz-basiert noch werturteil-begründet:** Das IQWiG stellt in seinem Methodenentwurf in Tabelle NT1 insgesamt acht Schwellenwerte als Grenzen für die Ableitung der Ausmaßkategorie eines Zusatznutzens in Abhängigkeit von der Qualität des Endpunkts und der Quantität des Effekts dar. Lediglich für einen dieser acht Schwellenwerte wurde vom IQWiG lediglich eine einzige Referenz angeführt, die die IQWiG-Annahme eines Effekts von ‚erheblichem' Ausmaß beim relativen Risiko für eine einzige Endpunkt-Kategorie (Gesamtmortalität) rechtfertigen soll. Alle weiteren sieben in Tabelle NT1 vom IQWiG für die übrigen Endpunktkategorien angeführten Schwellenwerte sind nicht werte-basiert und wurden vom IQWiG rechnerisch mittels einer mathematischen Rasterung „*als pragmatische Lösung*" ermittelt (vgl. S. 33 Anhang des IQWiG Methodenentwurfs). Die fehlende Begründung der mit der Herleitung der Schwellenwerte getroffenen quantitativen und qualitativen Wertbeurteilung ist nicht ausreichend angesichts der impliziten Tragweite der Schwellenwerte für Entscheidungen zur Patientenversorgung. Damit wird einer weiteren konkreten Kritik aus dem Stellungnahmeverfahren zur Nutzenbewertung nach § 35a SGB V von Ticagrelor keine Bedeutung beigemessen, da das IQWiG auch in seinem jetzigen Methodenentwurf wiederum nicht begründet, „*warum die Zielgröße ‚Mortalität' als grundsätzlich ‚bedeutender' eingestuft wird als die Zielgrößen ‚gesundheitsbezogene Lebensqualität' und/oder ‚schwerwiegende Symptome'.*" (Vgl. Stellungnahme Prof. Dr. Strech in Gemeinsamer Bundesausschuss: Zusammenfassende Dokumentation über die Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V Ticagrelor, vom 15. Dezember 2011 [13]).

Insofern ist ebenfalls kritisch zu sehen, dass die vom IQWiG vorgeschlagenen Kriterien zur Festlegung des Zusatznutzenausmaßes (Tabelle NT3) zwei in der AM-NutzenV explizit genannte Kriterien eines Zusatznutzens nicht konkret und direkt abbilden: ‚Heilung' und ‚spürbare Linderung der Erkrankung'. Gemäß § 5 Abs. 7, AM-NutzenV liegt ein erheblicher Zusatznutzen vor, wenn eine bisher nicht erreichte große Verbesserung des therapierelevanten Nutzens, insbesondere u.a. eine „*Heilung*" der Erkrankung erreicht wird, respektive ein beträchtlicher Zusatznutzen liegt vor, wenn eine bisher nicht erreichte

deutliche Verbesserung des therapierelevanten Nutzens erreicht wird, insbesondere u.a. eine „*spürbare Linderung der Erkrankung*“.

Hierzu führt das IQWiG im Methodenentwurf erläuternd aus, dass ‚Heilung‘ grundsätzlich einer Operationalisierung bedarf, „…*die sich regelhaft auf Kriterien stützen wird, die sich auch in den Endpunkten Mortalität und Morbidität abbilden lassen*“ und weiter, es sei für ‚Heilung‘ „…*die jeweilige konkrete Operationalisierung anhand der verwendeten Endpunkte daraufhin zu prüfen, ob sie einer relevanten Verbesserung der Mortalität bzw. schwerwiegender Ereignisse gleich kommt.“* Dabei konkretisiert das IQWiG jedoch nicht weiter, anhand welcher definitiven Kriterien geprüft und entschieden werden wird, ob und wann ‚Heilung‘ operationalisierende Endpunkte im Einzelfall eine relevante Verbesserung der Mortalität bzw. schwerwiegender Ereignisse abbilden. Die in diesem Zusammenhang getroffene Feststellung die „*Verkürzung der Symptomdauer, z.B. bei banalen Infektionskrankheiten, ist in diesem Sinne nicht als Heilung anzusehen*“ beinhaltet dagegen ein konkret getroffenes Werturteil, das jedoch vom IQWiG nicht begründet wird und ohne die Einbindung der Sichtweise von Betroffenen vom IQWiG nicht getroffen werden kann.

2.) **Die vom IQWiG vorgeschlagene Methodik zur Quantifizierung des Zusatznutzenausmaßes ist im Rahmen der Nutzenbewertung nach § 35a SGB V irrelevant, da sie vom Gemeinsamen Bundesausschuss bei seinen Entscheidung zum Ausmaß des Zusatznutzens keine Anerkennung findet:** Die vom IQWiG erneut im Methodenentwurf vorgeschlagene Methodik zur Ausmaßkategorisierung wurde vom IQWiG bereits im Rahmen der Dossierbewertung zu Ticagrelor im September 2011 erstmals vorgeschlagen und angewendet (Dossierbewertung A11-02, Seiten 86 - 92). Bereits in diesem Verfahren hat der Gemeinsame Bundesausschuss festgelegt, über den Nutzen nicht auf Basis dieser vom IQWiG angewendeten Methodik zu entscheiden und dies mit dem zusätzlichen kommentierenden Vermerk versehen, dass die Methodik zur Operationalisierung des Zusatznutzens Gegenstand weiterer Diskussionen sein wird. (Vgl. Gemeinsamer Bundesausschuss: Zusammenfassende Dokumentation über die Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V Ticagrelor, vom 15. Dezember 2011 [13]). Auch in allen seinen weiteren Verfahrensentscheidungen zur Nutzenbewertung nach § 35a SGB V hat der Gemeinsame Bundesausschuss bis heute auf diese vom IQWiG erstmals im Zuge des Bewertungsverfahrens zu Ticagrelor und jetzt erneut wieder im Methodenentwurf vorgeschlagene Methodik zur schwellenwertabhängigen Ausmaßkategorisierung nicht abgestellt, zuletzt wieder in seiner Entscheidung zu Pixantron vom 16. Mai 2013 [14]. Die methodische Vorgehensweise des IQWiG zur Ableitung des Ausmaßes des Zusatznutzens ist somit für die Verfahren zur Nutzenbewertung gemäß § 35a SGB V irrelevant, insofern da sie beim Gemeinsamen Bundesausschuss bei seinen diesbezüglichen Verfahrensentscheidungen nicht akzeptiert ist.

3.4 Vorschlag zur Ableitung der Gesamtaussage zum Zusatznutzen

In Bezug auf die zusammenfassende Bewertung zur Ableitung einer Gesamtaussage zum Zusatznutzen im Nutzenbewertungsverfahren gemäß § 35a SGB V (Schritt 3) beschreibt das IQWiG in Abschnitt 3.1.5 des Methodenentwurfs, dass die für jeden patientenrelevanten

Endpunkt einzeln getroffenen Aussagen zur Beleglage in einem bewertenden Fazit in Form einer Nutzen-Schaden-Abwägung – soweit möglich - zusammengefasst werden. Hierbei soll vom IQWiG anhand aller Endpunkte und unter Würdigung der Wahrscheinlichkeit und des Ausmaßes auf Endpunktebene die Gesamtaussage zum Zusatznutzen getroffen werden.

Jedoch wird das methodische Vorgehen für die Ableitung einer Gesamtaussage zum Zusatznutzen im Rahmen des Entwurfes nicht weiter ausgeführt. Es werden lediglich methodische Optionen genannt, ohne deren Relevanz für die Ableitung der Gesamtaussage aus methodischer Sicht zu würdigen. Bei der vom IQWiG im Methodenentwurf skizzierten Vorgehensweise, handelt es sich bei der Nutzen-Schaden-Abwägung um eine vom IQWiG durchgeführte Gewichtung von patientenrelevanten Endpunkten, die ohne Einbezug der Sichtweise von Betroffenen und ohne methodisch transparente Grundlage durchgeführt werden soll. Erforderlich ist jedoch eine Gewichtung von patientenrelevanten Endpunkten, die unter Einbezug der Sichtweise von Betroffenen und auf einer methodisch transparenten Grundlage durchgeführt wird. Unbedingt notwendig wäre an dieser für die Nutzenbewertung zentralen Stelle eine transparente Diskussion der methodischen Herangehensweise, wie patientenrelevante Endpunkte im Rahmen der Ableitung der Gesamtaussage zum Zusatznutzen gewichtet werden und in den Entscheidungsprozess einbezogen werden sollen.

In Abschnitt 3.1.5 des Methodenentwurfs nennt das IQWiG selbst zwei Möglichkeiten zur Durchführung einer Nutzen-Schaden-Abwägung, um zu einer Gesamtaussage zum Zusatznutzen zu kommen:

- Eine Möglichkeit der gleichzeitigen Würdigung von Nutzen und Schaden sei demnach „*die Gegenüberstellung der endpunktbezogenen Nutzen- und Schadenaspekte",* wobei „*die Effekte auf alle Endpunkte (qualitativ oder semiquantitativ) gegeneinander abgewogen werden mit dem Ziel, zu einer endpunktübergreifenden Aussage zum Nutzen bzw. Zusatznutzen einer Intervention zu kommen*" (vgl. S. 13 des Methodenentwurfs in der Version vom 18.04.2013).
- Die zweite Möglichkeit bestehe darin, „*die verschiedenen patientenrelevanten Endpunkte zu einem einzigen Maß zu aggregieren*" (vgl. S. 13 des Methodenentwurfs in der Version vom 18.04.2013). Hierbei werden die Aussagen zur Beleglage für jeden einzelnen patientenrelevanten Endpunkt gewichtet und beispielsweise in einem Summenscore zusammengefasst. Die international verbreiteten Nutzwerte für Gesundheitszustände werden vom IQWiG „*aufgrund der ethischen und methodischen Probleme*" (vgl. S. 13 des Methodenentwurfs in der Version vom 18.04.2013) abgelehnt. Stattdessen sollen „*alternative Verfahren der multikriteriellen Entscheidungsfindung oder der Präferenzerhebung*" (vgl. S. 13 des Methodenentwurfs in der Version vom 18.04.2013) angewendet werden. Hierzu werden zwei Methoden vorgeschlagen, allerdings ohne jegliche Einordnung und methodische Würdigung der Stärken und Schwächen.

Unklar bleibt an beiden Stellen jedoch, wie die Effekte (Nutzen und Schaden) genau gegeneinander abgewogen werden sollen bzw. welche Methodik angewendet werden sollte. Damit bleibt eine zentrale Fragestellung innerhalb der IQWiG-Methodik zur Nutzenbewertung, wie patientenrelevante Endpunkte zu einem aggregierten Nutzenmaß zusammengefasst werden sollen, weiterhin unbeantwortet.

Es muss hierbei kritisch hinterfragt werden, ob Ergebnisse, die auf einer unklaren methodischen Grundlage erzielt werden, so nachvollziehbar und reproduzierbar sein können, damit sie den Transparenzerfordernissen des IQWiG und eines Verfahrens zur Nutzenbewertung gemäß § 35a SGB V gerecht werden. Zu fordern ist daher zunächst eine transparente Diskussion der methodischen Grundlage der zusammenfassenden Bewertung von Nutzen- und Schadenparametern.

Das IQWiG selbst fordert, dass die gleichzeitige Würdigung von Nutzen und Schaden prospektiv im Berichtsplan oder im Vorbericht beschrieben werden sollte. Im Fall der quantitativen Gewichtung von patientenrelevanten Endpunkten mittels Summenscores sollte die Nutzen-Schaden-Abwägung prospektiv „*zum Zeitpunkt der Auswahl der zu untersuchenden Endpunkte erfolgen*" (vgl. S. 13 des Methodenentwurfs in der Version vom 18.04.2013). Bei den bislang im Rahmen der Nutzenbewertungsverfahren gemäß § 35a SGB V erfolgten zusammenfassenden Bewertungen zur Ableitung einer Gesamtaussage zum Zusatznutzen durch das IQWiG ist eine derartige prospektive Beschreibung der gleichzeitigen Würdigung von Nutzen- und Schadenparametern jedoch nicht erfolgt.

Aus Sicht der Novartis Pharma GmbH ist eine Ableitung der Gesamtaussage zum Zusatznutzen auf transparenter methodischer Grundlage zu fordern.

Für die beiden vom IQWiG zur Aggregation multipler patientenrelevanter Endpunkte vorgeschlagenen Verfahren der multikriteriellen Entscheidungsfindung oder der Präferenzerhebung fehlt eine umfassende Würdigung der methodischen Stärken und Schwächen beider Ansätze. Es kann angenommen werden, dass weder der Analytic Hierarchy Process noch die Conjoint Analyse alle Anforderungen erfüllen, die an eine Methode zur Gewichtung multipler patientenrelevanter Endpunkte zu stellen wären [15]. Während der Analytic Hierarchy Process unter Praktikabilitätsaspekten entscheidende Vorteile aufweist, stellt die (Choice-Based) Conjoint Analyse das aus nutzentheoretischer Sicht anzuwendende Verfahren dar.

Aus Sicht der Novartis Pharma GmbH bedarf es an dieser Stelle einer umfassenden und transparenten Diskussion mit der Fachöffentlichkeit zum methodischen Vorgehen bei der Endpunktgewichtung. Die Ergebnisse der vom IQWiG finanzierten Pilotstudien zu beiden Verfahren sollten unverzüglich veröffentlicht werden. Die methodischen Vor- und Nachteile sollten transparent im Rahmen eines Stellungnahmeverfahrens diskutiert werden. Mögliche Alternativvorschläge zur Nutzen-Schaden-Abwägung sollten zugelassen werden. Die zusammenfassende Bewertung zur Ableitung einer Gesamtaussage zum Zusatznutzen muss auf Basis einer transparenten Methodik erfolgen, die vor ihrer praktischen Anwendung zur Entscheidungsunterstützung einer umfassenden Expertendiskussion unterzogen werden sollte.

Anstelle einer Notwendigkeit für eine umfassende Methodendiskussion, wie patientenrelevante Endpunkte gewichtet werden sollen, formuliert das IQWiG abschließend auf Seite 19 des jetzt vorgelegten Methodenentwurfs: *„Für den dritten Schritt der Operationalisierung, der Gesamtaussage zum Ausmaß des Zusatznutzens bei gemeinsamer Betrachtung aller Endpunkte, ist eine strenge Formalisierung nicht möglich, da für die hierzu zu treffenden Werturteile gegenwärtig keine ausreichende Abstraktion bekannt ist. Das Institut wird im Rahmen seiner Nutzenbewertung die Aussagen zur Wahrscheinlichkeit und*

zum Ausmaß der Effekte vergleichend gegenüberstellen und einen begründeten Vorschlag für eine Gesamtaussage unterbreiten."

Fraglich ist inwiefern und auf welcher Grundlage das IQWiG - als wissenschaftlich unabhängiges und nicht demokratisch legitimiertes Institut - eine derartige wertbasierte Entscheidung herbeiführen kann, indem es einen – wie auch immer - begründeten Vorschlag für eine Gesamtaussage und damit zur Gewichtung von patientenrelevanten Endpunkten vorlegt – ohne diesen Vorschlag auf Basis einer anerkannten methodischen Grundlage zu erarbeiten. Die Ableitung einer Gesamtaussage zum Zusatznutzen einer medizinischen Intervention im Vergleich zur zweckmäßigen Vergleichstherapie auf Basis einer endpunktspezifischen Bewertung des medizinischen Zusatznutzens unter Abwägung von Nutzen- und Schadenaspekten beinhaltet Werturteile, die definitionsgemäß von einem unabhängigen wissenschaftlichen Institut ohne Einbeziehung der Sichtweise von Betroffenen nicht getroffen werden kann.

Aus Sicht der Novartis Pharma GmbH ist die Anwendung der vom IQWiG vorgeschlagenen Verfahren der multikriteriellen Entscheidungsfindung oder der Präferenzerhebung bei der Bestimmung des Ausmaßes an Zusatznutzen im Rahmen der Nutzenbewertung gemäß § 35a SGB V zunächst in Bezug auf deren Verwertbarkeit für die Priorisierung und Gewichtung multipler patientenrelevanter Endpunkte öffentlich zu diskutieren.

Eine Darlegung des allgemeinen methodischen Vorgehens zur Gewichtung multipler patientenrelevanter Endpunkte sollte indikationsunabhängig im Rahmen des IQWiG-Methodenpapiers thematisiert werden. Ohne festgelegte Methodik zur Vorgehensweise besteht die Gefahr der willkürlichen und nicht entscheidungskongruenten Bewertung im Rahmen einzelner Nutzenbewertungsverfahren. Im jetzigen Methodenentwurf des IQWiG fehlen Hinweise zur weiteren Klärung der noch offenen methodischen Fragen. Alternativen werden nicht aufgezeigt. Die methodische Schlussfolgerung aus Kapitel 3.1.5, wie bei der Endpunktgewichtung weiter vorgegangen werden soll und welche Relevanz die beiden vorgestellten Methoden im Rahmen der Nutzenbewertung gemäß § 35a SGB V haben, bleibt damit im Methodenentwurf weithin unklar.

Referenzen

1. European Medicines Agency. Points to consider on application with: 1. meta-analyses; 2. one pivotal study [online]. 31.05.2001 [Zugriff: 22.09.2010]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf.
2. Gemeinsamer Bundesausschuss. Geschäftsordnung und Verfahrensordnung. Zuletzt geändert 21.02.2013. BAnz AT 15.05.201. In Kraft getreten am: 16.05.2013
3. Gemeinsamer Bundesausschuss. Beschluss über eine Änderung der Verfahrensordnung: Änderungen im 5. Kapitel – Neufassung der Modulvorlagen in der Anlage II. Beschluss vom 18. April 2013.
4. Gemeinsamer Bundesausschuss. Beschluss über eine Änderung der Verfahrensordnung: Änderungen im 5. Kapitel – Neufassung der Modulvorlagen in der Anlage II. Beschluss vom 18. April 2013. Änderungen der Modulvorlagen zu Modul 3.
5. Spitzenverband Bund der Krankenkassen und Deutschen Apothekerverband e. V. Rahmenvertrag über die Arzneimittelversorgung nach § 129 Absatz 2 SGB V in der Fassung vom 15. Juni 2012. http://www.gkv-Spitzenverband.de/media/dokumente/krankenversicherung_1/arzneimittel/rahmenvertraege/apotheken/AM_20120615_S_RVtg_129_Abs2.pdf. (Zuletzt geprüft 21.05.2013)
6. Mauskopf JA, Sullivan SD, Annemans L, et al. Principles of Good Practice for Budget Impact Analysis: Report of the ISPOR Task Force on Good Research Practices – Budget Impact Analysis. *Value in Health* 2007;10:336-47.
7. Budget Impact Analysis Good Practices II: A Report of the ISPOR Budget Impact Good Practices Task Force DRAFT Version 29 April 2013. http://www.ispor.org/TaskForces/documents/ISPOR_BIA_Report_Draft_%20April_29_2013.pdf
8. Graf von Schulenburg JM et al. Deutsche Empfehlungen zur gesundheitsökonomischen Evaluation – dritte und aktualisierte Fassung des Hannoveraner Konsens. Gesundh ökon Qual manag 2007; 12:285-90.
9. Graf von der Schulenburg JM, Vauth C, Mittendorf T, Greiner W (2007): Methoden zur Ermittlung von Kosten-Nutzen-Relationen für Arzneimittel in Deutschland, In: Gesundheitsökonomie & Qualitätsmanagement 2007;12:3-25.
10. Braun S, Prenzler A, Mittendorf T, Graf von der Schulenburg JM (2009): Bewertung von Ressourcenverbräuchen im deutschen Gesundheitswesen aus Sicht der Gesetzlichen Krankenversicherung, In: Gesundheitswesen 2009; 71:19-23.
11. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. IQWiG-Berichte – Jahr 2011 Nr. 96 Ticagrelor – Nutzenbewertung gemäß § 35a SGB V. http://www.g-ba.de/informationen/nutzenbewertung/
12. Gemeinsamer Bundesausschuss. Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über die Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V - Ticagrelor vom 15. Dezember 2011. http://www.g-ba.de/informationen/nutzenbewertung/
13. Gemeinsamer Bundesausschuss. Zusammenfassende Dokumentation über die Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V- Ticagrelor. 15. Dezember 2011. http://www.g-ba.de/informationen/nutzenbewertung/
14. Gemeinsamer Bundesausschuss. Tragende Gründe zum Beschluss des Gemeinsamen Bundesaus-schusses über eine Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII - Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V – Pixantron vom 16. Mai 2013. http://www.g-ba.de/informationen/nutzenbewertung/
15. Neidhardt K, Wasmuth T, Schmid A: Die Gewichtung multipler patientenrelevanter Endpunkte – Ein methodischer Vergleich von Conjoint Analyse und Analytic Hierarchy Process unter Berücksichtigung des Effizienzgrenzenkonzepts des IQWiG, Diskussionspapier 02-12, Universität Bayreuth, Link: www.fiwi.uni-bayreuth.de/de/download/WP_02-12.pdf

A.1.21 – Pfizer Deutschland GmbH

Stellungnahme der Pfizer Deutschland GmbH zum Entwurf des IQWiG

„Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“ vom 18.04.2013

Die Pfizer Deutschland GmbH möchte positiv hervorheben, dass das IQWiG sich dazu entschieden hat, seine Änderungen am Methodenpapier (Version 4.0) zur wissenschaftlichen Diskussion in Form eines Stellungnahmeprozesses zu öffnen. Daher möchten wir uns gerne an der Diskussion beteiligen und reichen folgende Stellungnahme ein.

Im Rahmen der Stellungnahme werden dabei folgende Punkte kommentiert :

- Umgang mit Heterogentitäten im Rahmen von Meta-Analysen
- Festlegung der Schwellenwerte für das Ausmaß des Zusatznutzens anhand der „Matrix“
- Interaktionstests bei Subgruppenanalysen
- Übertragbarkeit
- Saldierung von Nutzen und Schaden

1 Umgang mit Heterogentitäten im Rahmen von Meta-Analysen

Das IQWiG beschreibt in Abschnitt 7.3.8. den Umgang bei der Beurteilung von Ergebnissicherheiten im Fall einer heterogenen Studienlage, insb. die Einführung des Prädiktionsintervalls.

- Generell stimmen wir damit überein, dass ein Effektschätzer aus MA bei heterogener Studienlage nicht (so) zuverlässig ist; Methoden, die helfen Aussagen darüber zu treffen, wie der Effekt in dieser heterogenen Situation aussehen könnte, sind wünschenswert.

- Allerdings möchten wir anmerken, dass das Prädiktionsintervall ist nicht hinreichend untersucht ist. Aus diesem Grund lehnt das IQWiG zum Beispiel bei der klassischen NB nach §35b SGB V die Bewertung unter Hinzunahme von indirekten Vergleichen regelhaft ab.

    - 2006 erstmals als Poster vorgestellte Methode
    - Erstmalig in peer reviewed Literatur beschrieben 2011 (Riley 2011)
    - In der praktischen Anwendung, insbesondere bei vergleichbaren Instituten internationaler Entscheidungsträger, (noch) nicht (wirklich) eingesetzt, daher kaum Erfahrungen mit dem Intervall
    - Keine hinreichende Untersuchung der Eigenschaften

- Das IQWiG berechnet (i.d.R.) das Prädiktionsintervall (PI) erst ab 4 Studien und begründet das mit der damit verbundenen fehlenden Präzisierung der Schätzung.
  Ein Effekt kann dabei nur deutlich gleichgerichtet sein, wenn das PI den Nulleffekt nicht überdeckt. Ansonsten gilt eine Regel (s.u.), wann die Effekte zumindest mäßig gleichgerichtet sind.
  Für weniger als 4 Studien (wenn kein PI berechnet wird), gelten diese Regeln ebenfalls, allerdings ist es möglich hier zum Schluss von deutlich gleichgerichteten Effekten zu kommen. Ist diese Regel konsistent mit der Regel für PI?

- Die Regel, dass bei gleichgerichteten Studien mind. 80% der Studiengewichte (im Random Effect Model, REM) gleichgerichtet sein müssen und mind. 50% der Studiengewichte aus statistisch signifikanten Studien kommen müssen, ist nicht mit Evidenz belegt. Wir gehen daher davon aus, dass diese Grenzen nicht begründet werden können. Darüber hinaus liegen

unseres Wissens keine wissenschaftlichen Untersuchungen über die Konsequenzen dieser Grenzen bei unterschiedlicher Ausgangssituation vor.

- Durch die Einführung der Kategorien „deutlich“ und „mäßig“ gleichgerichtet hat sich eine Verschärfung der Anforderungen der Evidenz an die Ergebnissicherheit ergeben. Es ist nicht begründet, warum das IQWiG von der Festlegung der Beleglage auf Basis von „gleichgerichteten Effekten“ aus dem Methodenpapier 4.0 abgewichen ist und diese Verschärfung eingeführt wurde.

- Um das Prädiktionsintervall zu berechnen und graphisch nach den Vorstellungen des IQWiG darzustellen, bedarf es einer separaten software oder einer Anpassung der existierenden. Beides ist bislang nicht marktreif verfügbar. Allerdings ist es erforderlich, diese Analysen bereits bei der Erstellung des Dossiers zur frühen Nutzenbewertung darzustellen und entsprechend zu diskutieren, um zu der nach den Methoden des IQWiG adäquaten Bewertung der Ergebnissicherheit zu kommen.

2 Festlegung der Schwellenwerte für das Ausmaß des Zusatznutzens anhand der „Matrix“

Für die Bestimmung des Ausmaßes des Zusatznutzens hat das IQWiG im Rahmen seiner ersten Bewertung eines Dossiers zur frühen Nutzenbewertung (Ticagrelor) ein Klassifikationsschema entwickelt, das auf Konfidenzintervallen von relativen Risiken beruht.
Es ist positiv hervorzuheben, dass ein transparentes und objektives Verfahren angestrebt wurde das die Größe des Zusatznutzens aufgrund der Daten bestimmt. Es ist aber aus Sicht von Pfizer auch wichtig, dass das Verfahren praktikabel und planbar bleibt. Weiter ist positiv hervorzuheben, dass bei den Analysen von Raten die relativen Maße, wie relatives Risiko und Hazard Ratio die Grundlage der Konfidenzintervalle bilden.

Dennoch möchten wir im Rahmen dieser Stellungnahme einige wesentliche und aus unserer Sicht überarbeitungswürdige Punkte ansprechen:

- Der Ausgangswert für die Entwicklung der Konfidenzintervalle für die quantitative Bewertung des Zusatznutzens (erheblich/beträchtlich/gering) beruht auf einer einzigen Publikation (Djulbegovic, 2008), die sich wiederum 1) auf Aussagen der Autoren und 2) ausschließlich auf die Indikation der Onkologie bezieht. Aus dieser Publikation leiten die Autoren des Methodenpapiers ab, dass nur Substanzen mit einem Effekt auf die Mortalität (Relatives Risiko oder Hazard Ratio) von mindestens 0.5 und dem entsprechenden Konfidenzintervall einen erheblichen Zusatznutzen haben. Es darf in Frage gestellt werden, ob dieses „one size fits all“-Prinzip des Ansatzes den Auffassungen der Patienten gerecht wird. Zudem stellt sich die Frage, ob eine solche normative Festlegung keiner entsprechenden Legitimation bedarf. Pfizer schlägt daher vor, dass sich das IQWiG im Vorfeld der Bewertung, optimal im Vorfeld der frühen Beratungsgespräche mit dem G-BA, mit Patienten, Experten und Herstellern indikationsspezifische Schwellenwerte erarbeitet.

- Die Ableitung der zu erreichenden Effektschätzer für das Ausmaß des Zusatznutzens, ausgehend von oben erwähntem Relativem Risiko (Hazard Ratio) von 0.5 für erheblichen Zusatznutzen beim Endpunkt Mortalität, wird beschrieben als mathematische Regel. „Dabei war zu beachten, dass die Anforderungen von der Zielgrößenkategorie „Mortalität“ ausgehend für weniger schwerwiegende Zielgrößen zunehmen und von der Ausmaßkategorie „erheblich“ ausgehend für niedrigere Ausmaßkategorien abnehmen sollten. Eine Rasterung von 1/6 für die tatsächlichen Effekte erwies sich dabei als pragmatische Lösung. Nachfolgend werden die Schwellenwerte für die jeweiligen Ausmaßkategorien beschrieben.“ Auch hier darf zu Bedenken gegeben werden, ob nicht die

Beteiligung von betroffenen Patienten sowie die Berücksichtigung der zugrundeliegenden Indikation adäquater – wenn auch zugegebener maßen nicht pragmatischer – angeratener ist.

- Die Ableitung der Grenzen des Konfidenzintervalls wurde unter der Annahme von zwei gut geplanten und auf den entsprechenden Zielparameter adäquat gepowerten Studien durchgeführt. Die Autoren zeigen, dass sich unter dieser Annahme bei Anwendung der oberen Konfidenzgrenzen als Schwellenwerte kein Verlust an Power ergibt. Eine Auswertung der ersten 26 Bewertungen durch das IQWiG ergab, dass die Bewertung in 9 Fällen ohne zugrunde liegende Daten, in 8 Fällen auf Basis einer einzigen Studie, in 6 Fällen auf 2 Studien und in 3 Fällen auf 3 Studien basierte (Quelle: interne Informationen des vfa). Diese Untersuchung kommt zu dem Ergebnis, dass in ca. 30 Prozent nur eine Studie herangezogen werden kann. Durch die Zulassungsanforderungen insbesondere bei onkologischen Präparaten kann davon ausgegangen werden, dass sich dieses Bild gerade bei Onkologika auch in der Zukunft nicht verändern wird. Zudem wird das Anwendungsgebiet durch den gemeinsamen Bundesausschuss (G-BA) bei der Wahl der zweckmäßigen Vergleichstherapie häufig aufgeteilt (z.B. in einzelne Studien, siehe Bsp. Apixaban) die Studien aufgeteilt, so dass für die endgültige Bewertung nicht die gesamte Patientenzahl, und damit auch nicht mehr die ursprünglich geplante Power, zur Verfügung steht. Hier sind neue Ansätze wünschenswert, die die Problematik reflektieren und insbesondere die Power der Studien auch in der Realität erhalten. Darüber hinaus soll angemerkt werden, dass die Möglichkeit nur einer pivotalen Studie als Ausnahme für die Ableitung eines Nutzenbelegs zwar in den Änderungen erwähnt wird („Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und ihre Ergebnisse besondere Anforderungen zu stellen“, S. 10), diese jedoch bei der Entwicklung der Konfidenzgrenzen nicht berücksichtigt wird. Die Anzahl der Studien fließt damit nicht nur in die Bewertung der Ergebnissicherheit des Zusatznutzens sondern auch in dessen Ausmaß ein und wird somit doppelt gezählt.

- Für ordinale und stetige Variablen wird kein Schwellenwert zur Bestimmung des Ausmaßes des Zusatznutzens bestimmt. Vielmehr wird gefordert, die stetigen/ordinalen Variablen zu dichotomisieren und damit eine Berechnung der geforderten Kenngröße (Relatives Risiko) zu ermöglichen. Die Abweichung von der aktuellen Studienplanung mit stetigen Größen wird dabei zu (teilweise nicht unerheblichem) Verlust der Power führen. Darüber hinaus kann es auch durchaus Situationen geben, in denen keine allgemein akzeptierten Responder-Kriterien für die entsprechende Variable existieren. Nach aktuell vorgeschlagener Methode kann für diese Variablen ausschließlich aufgrund der vom IQWiG gewählten Methode kein Zusatznutzen festgestellt werden. Es bleibt unklar, warum gut untersuchte Methoden zur Bestimmung der klinischen Relevanz, wie sie das IQWiG zum Teil in früheren Vorträgen auch selbst vorgestellt und propagiert hat, zumindest in solchen Situationen nicht zum Einsatz kommen.

3 Interaktionstests bei Subgruppenanalysen

Das IQWiG testet – analog zur Untersuchung von Heterogenität bei Meta-Analysen - in Einzelstudien auf Subgruppeneffekte. Dabei verwendet es Interaktionstests und spricht es bei einem Signifikanzwert von $p<0,05$ von einem Beleg, bei einem Niveau von $0,05 < p < 0,2$ von einem Hinweis und bei $p>0,2$ von Fehlen eines Subgruppeneffekts.

- Hierbei möchten wir zu bedenken geben, dass durch die inflationäre Anzahl von Tests auf Subgruppeneffekte das Risiko einer irrtümlichen Signifikanz unkontrollierbar steigt, da durch die explorative Fragestellung der alpha-Fehler nicht weiter kontrolliert wird. Dennoch wird aus den Ergebnissen dieser Tests der „Beleg“ eines Subgruppeneffekts abgeleitet und damit eventuell der Zusatznutzen einer neuen Substanz in gewissen Patientenpopulationen

fälschlich verneint. Wir sind uns der Tatsache bewusst, dass auch Zulassungsbehörden post hoc Subgruppenanalysen betrachten, möchten aber dennoch ein Wort der Vorsicht bei der Interpretation und auch Anzahl von geforderten und durchgeführten – nicht a priori geplanten – Subgruppenanalysen aussprechen.

- Das Institut verwendet darüber hinaus dieselbe Wortwahl für Subgruppentests wie für die Ergebnissicherheit bei der Abschätzung des Ausmaßes des Zusatznutzens. Während diese Differenzierung den statistisch belesenen Fachleuten verständlich ist, haben Erfahrungen in der Diskussion mit fachfremden Entscheidungsträgern (z.B. im Rahmen von Anhörungen) gezeigt, dass die Analogie der Begriffe zu einer Gleichsetzung der Bedeutung führen kann. Um Missverständnisse dieser oder ähnlicher Art künftig zu vermeiden schlägt Pfizer vor, für die Ergebnisse der Interaktionstests für eine deutlichere Differenzierung ein anderes wording zu verwenden.

4 Übertragbarkeit

Das IQWiG verlangt für die Bewertung des Zusatznutzens eine Übertragbarkeit der vorliegenden Evidenz auf die deutsche Behandlungssituation. Hinsichtlich der Patientenpopulation präzisiert es dies wie folgt:

„Für das Einschlusskriterium bezüglich Population reicht es aus, wenn bei mindestens 80 % der in der Studie eingeschlossenen Patienten dieses Kriterium erfüllt ist. Liegen für solche Studien entsprechende Subgruppenanalysen vor, wird auf diese Analysen zurückgegriffen. Studien, bei denen das Einschlusskriterium bezüglich Population bei weniger als 80 % der in der Studie eingeschlossenen Patienten erfüllt ist, werden nur dann eingeschlossen, wenn entsprechende Subgruppenanalysen vorliegen. Ist beides in einer Studie nicht der Fall, wird die Studie aus der Nutzenbewertung ausgeschlossen."

- In einigen Fällen wird von der Zulassungsbehörde aufgrund von den Daten der klinischen Prüfung die Indikation eingeschränkt (z.B. Therapielinie, Schweregrad der Erkrankung,…). Entsprechende Subgruppenauswertungen für den zu bewertenden neuen Wirkstoff sind zwar theoretisch durch den Patentinhaber möglich, allerdings auf Kosten der Power der Studie (zu den Konsequenzen des Powerverlusts siehe auch 2)
- In der Praxis werden allerdings auch – wenn auch eher seltener - die Studienergebnisse der untersuchten Patientenpopulation auf eine andere – nicht untersuchte Population interpoliert und die Substanz dann (ausschließlich) auf diese Population zugelassen. Der Hersteller, der die Zulassung beantragt hat, hat auf diesen Prozess nur sehr eingeschränkt Einfluss. Für die anstehende Nutzenbewertung resultiert das Vorgehen der Zulassungsbehörden allerdings in einem völligen Fehlen von Evidenz (in der gefragten Population). Eine Aussage zum Zusatznutzen kann daher rein formell nicht mehr getroffen werden.
- Pfizer schlägt daher vor, in den Änderungen des Methodenpapiers Verfahren zur Interpolation analog denen der Zulassungsbehörden zu akzeptieren. Auch die Verfahren des GRADEings, die die internationalen Standards der Evidenzbasierten Medizin (EbM) beschreiben, sollten hierbei zumindest in Betracht gezogen werden.

5 Saldierung von Nutzen und Schaden

Das IQWiG schlägt für die „klassische Nutzenbewertung" nach §35b Methoden für eine zusammenfassende Bewertung von Nutzen und Schaden vor (Abschnitt 5.1.3).

- Die gleiche Problematik der „Saldierung" ergibt sich allerdings auch in der Bewertung der Nutzendossiers nach §35a. Wenn auch für einen verpflichtend Einsatz bei den frühen Nutzenbewertungen sicherlich weitere Forschungen bzw. Anpassungen notwendig sind, so sollte doch geprüft werden, in wieweit diese Verfahren - insbesondere die Conjoint analyse unter Einbeziehung der Patienten - für die Saldierung auch in der frühen Nutzenbewertung anwendbar sind.

Referenzen:

Djulbegovic et al, 2008; Treatment Success in Cancer; Arch Intern Med.; 168(6):632-642

Riley et al, 2011; Interpretation of random effects meta-analyses; BMJ; 342:d549

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) (2013); „Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“ vom 18.04.2013; abrufbar unter https://www.iqwig.de/de/methoden/methodenpapiere/allgemeine_methoden.3020.html

A.1.22 – Sanofi-Aventis Deutschland GmbH

Stellungnahme der Sanofi-Aventis Deutschland GmbH zum IQWiG Methodenpapier Version 4.1.

Zum Stellungnahmeverfahren allgemein:

Im IQWiG- Entwurf vom 18.04.2013 zur „Aktualisierung Allgemeine Methoden“ wird das Procedere des Stellungnahmeverfahrens wie folgt beschrieben: „Im Anschluss an das Stellungnahmeverfahren werden die betreffenden Abschnitte ggf. überarbeitet und in die Allgemeinen Methoden integriert. Hieraus gehen dann die Allgemeinen Methoden Version 4.1. hervor.“ (IQWiG 2013) Auf der Homepage des IQWiG wird ersichtlich, dass es sich um ein schriftliches Stellungnahmeverfahren handelt. (IQWiG 2013a) Wesentliche Bestandteile der Aktualisierung befassen sich mit den Themen „Anforderung an die Beleglage für Aussagen zum Nutzen mit unterschiedlichen Aussagesicherheiten“ und „Operationalisierung des Ausmaßes des Zusatznutzens“. (IQWiG 2013) Diese Methodik wurde erstmals beim Verfahren der Nutzenbewertung gemäß §35a SGB V von Ticagrelor angewendet und in der Folge vielfach bei allen weiteren Nutzenbewertungen. Damals wurde im Rahmen der mündlichen Anhörung zugesichert, dass der G-BA die Methodik der Kategorisierung des Zusatznutzens zu einem späteren Zeitpunkt zusammen mit Experten und dem IQWiG in einer Veranstaltung diskutieren würde. (G-BA. Wortprotokoll Anhörung Ticagrelor) Diese Veranstaltung fand mit großer zeitlicher Verzögerung und in sehr kleinem Rahmen statt.

Im Sinne der Pflicht des IQWiG gemäß § 139a SGB V, in regelmäßigen Abständen über Arbeitsprozesse und – ergebnisse einschließlich der Grundlagen für die Entscheidungsfindung öffentlich zu berichten, und im Sinne eines transparenten wissenschaftlichen Austausches auf Augenhöhe wäre eine mündliche Anhörung im Nachgang zum schriftlichen Stellungnahmeverfahren unerlässlich. Diese Möglichkeit eines offenen, konstruktiven und transparenten Austausches sollte das IQWiG im Sinne zukünftiger Nutzenbewertungen und letztlich im Sinne der Patienten nicht ungenutzt verstreichen lassen.

Zu 2.2.3. Review der Produkte des Instituts

Im letzten Absatz legt das Institut dar, dass ihm, neben dem externen Qualitätssicherungsverfahren unter Beteiligung vom Institut ausgewählter und beauftragter Reviewer, auch ein offenes und unabhängiges Reviewverfahren nach Veröffentlichung der Institutsprodukte wichtig ist. Für dieses öffentliche Reviewverfahren wäre es außerordentlich wichtig, die Meinungen und Einschätzungen der im Rahmen des Qualitätssicherungsverfahrens beauftragten Reviewer zu kennen und ggf. auch die Gründe des Instituts, wenn vom Rat der Reviewer abgewichen wurde. Die Position des Instituts, die extern beauftragten Reviews nicht zu veröffentlichen, bedeutet eine Einschränkung der Transparenz des Prozesses, in dem das jeweilige Institutsprodukt zustande kommt.

Zu 3.1.4. Endpunktbezogene Bewertung

Das Methodenpapier des IQWiG legt fest, dass die Belegbarkeit des Zusatznutzens oder –schadens in den 4 Kategorien „Beleg“, „Hinweis“, „Anhaltspunkt“ oder „keines davon“ erfolgt.

(IQWiG 2013) Grundsätzlich muss festgestellt werden, dass diese Kategorisierung weder hergeleitet und nachvollziehbar begründet wird, noch international angewendeten HTA-Standards entspricht (wie es der Gesetzgeber in §139a SGB V Abs. 4 explizit fordert). International anerkannt und üblich ist dagegen eine Darstellung der Güte von Studien, z.B. anhand der GRADE-Kriterien. (Grade Working Group 2004))

Auch die vom IQWiG verwendete Kategorisierung zum Ausmaß der qualitativen Ergebnissicherheit ist nicht eindeutig nachvollziehbar, da es keine Operationalisierung z.B. zur Unterscheidung von hohem oder niedrigem Verzerrungspotential gibt und die Gesamteinschätzung dadurch nicht immer nachvollziehbar und transparent ist - und unter Umständen sogar subjektiv gefärbt ist.

Zur Ableitung eines „Belegs für Zusatznutzen“ fordert das IQWiG mindestens 2 Studien mit statistisch signifikantem Ergebnis der Metaanalyse bei vorhandener Homogenität. Zur Ableitung der Beleglage bei heterogenen Ergebnissen aus Meta-Analysen fordert das IQWiG gleichgerichtete Effekte in Studien mit den Bedingungen: das Gesamtgewicht dieser „gleichgerichteten“ Studien soll ≥ 80% sein und mindestens 2 dieser Studien sollen statistisch signifikante Ergebnisse zeigen. (IQWiG 2013) An dieser Stelle fehlt eine Begründung für die Festlegung des Gesamtgewichtes auf mindestens 80% für die Feststellung gleichgerichteter Effekte.

Sofern es sich nicht um primäre Endpunkte der betrachteten Studien handelt, war die Fallzahlplanung der Studien nicht auf das Erreichen statistischer Signifikanz ausgelegt. In dieser Situation ist Signifikanz, wenn überhaupt, nur in einer Meta-Analyse nachweisbar und die Forderung, dass zwei Studien einzeln signifikant sein müssen, nicht zu begründen.

Für den Fall von weniger Studien oder schwächerer Ergebnissicherheit dieser Studien behält sich das IQWiG die Kategorien „Hinweis“ und „Anhaltspunkt“ vor. Bei einer einzelnen Studie mit hoher qualitativer Ergebnissicherheit wird dementsprechend kein „Beleg“ mehr, sondern nur noch ein „Hinweis“ zuerkannt. (IQWiG 2013) Das IQWiG stellt in Ausnahmefällen die Möglichkeit in Aussicht, auch bei Vorliegen einer einzelnen Studie einen „Beleg“ zu erhalten, wenn bestimmte Voraussetzungen erfüllt sind, die in den „Points to consider on application with 1. meta-analyses; 2. one pivotal study“ der European Medicines Agency, der europäischen zentralen Zulassungsbehörde, näher ausgeführt sind. Dies ist sehr zu begrüßen, da gerade für Patienten, für die sonst keine entsprechenden Therapiealternativen mehr vorhanden sind, aus ethischen Gründen auch keine neuerlichen randomisierten Studien mit einem Medikament möglich sind, das seine Überlegenheit bereits unter Beweis gestellt hat. Dies ist häufig der Fall bei onkologischen Erkrankungen, wo es unethisch wäre, das einzige noch wirkende Produkt den Patienten im Kontrollarme vorzuenthalten. Schon aus diesem Grund würde eine entsprechende zweite Studie durch Ethikkommissionen nicht genehmigt werden, weshalb die EMA in diesen Fällen bereits bei Vorliegen von nur einer Phase III Studie in Anwendung des oben zitierten Dokuments eine Zulassung erteilt. In der Vergangenheit musste jedoch leider festgestellt werden, dass das IQWiG an dieser Stelle von seiner eigenen Methodik Abstand nimmt.

Zusätzlich sollte berücksichtigt werden, dass der Anteil der in eine Studie eingeschlossenen Patienten, bezogen auf die Gesamtpopulation der an einer bestimmten Krankheit leidenden Patienten, in einer einzelnen Studie bei einer selteneren Erkrankung oftmals höher ist als bei einer anderen, häufiger auftretenden Indikation die Patientenanteile aller Phase III-Zulassungsstudien gemeinsam. Eine zu starre Fokussierung auf die bloße Anzahl der

Studien, die zur Zulassung geführt haben, ohne Berücksichtigung der Repräsentativität, benachteiligt somit unter Umständen Patienten, die an einer selteneren Erkrankung leiden.

In Tabelle 2, S. 11, wird die Beleglage für unterschiedliche Aussagesicherheiten nochmals tabellarisch dargestellt. (IQWiG 2013) In vorangehenden Diskussionen mit dem IQWIG wurde die Frage aufgeworfen, ob bei einer einzelnen Studie mit direktem Vergleich gegenüber der zweckmäßigen Vergleichstherapie die Beleglage durch die Hinzunahme indirekter Vergleiche aufgewertet werden kann, um damit eine höhere Stufe der Aussagesicherheit zu erreichen. Insbesondere Herr Prof. Windeler schloss diese Möglichkeit nicht aus. (IV. Expertenforum Pharma, 25.10.2012) Wir schlagen daher vor, Tabelle 2 um Szenarien zu ergänzen, in denen die Aufwertung der Aussagesicherheit unter zusätzlicher Berücksichtigung indirekter Vergleiche abgebildet ist.

Indirekte Vergleiche können darüber hinaus, sofern sie mit direkten Vergleichen konsistent sind, auch hilfreich sein, die Präzision der Schätzung von Abständen zur zweckmäßigen Vergleichstherapie zu verbessern.

Im Folgenden wird auf Seite 12 präzisiert, dass „die für einen Beleg notwendige Bestätigung (…) eines statistisch signifikanten Ergebnisses einer Studie hoher qualitativer Ergebnissicherheit (…) durch (…) Ergebnisse mäßiger (…) qualitativer Ergebnissicherheit erbracht werden“ können. (IQWiG 2013) An dieser Stelle fehlt eine entsprechende Rationale dafür, warum der Bereich für das Gewicht einer Studie hoher qualitativer Ergebnissicherheit auf 25% bis 75% festgelegt wird.

Zu 3.1.5. Zusammenfassende Bewertung

Dieses Kapitel behandelt eher theoretisch, welche Möglichkeiten das IQWiG sieht, die verschiedenen Nutzen- und Schadensaspekte innerhalb eines Maßes zu aggregieren, z. B. durch die Gewichtung patientenrelevanter Endpunkte zur Bildung eines Summenscores. (IQWiG 2013) Eine detailliertere Ausführung fehlt jedoch. Es wird betont, dass die Methodik der Aggregation möglichst in einem Berichtsplan oder Vorbericht spezifiziert werden sollte. Da der Prozess der frühen Nutzenbewertung gem. §35a SGB V solch einen Berichtsplan bzw. Vorbericht nicht vorsieht, ist weiterhin unklar, wie in den Verfahren zur frühen Nutzenbewertung vorgegangen wird. Im Sinne einer Planungssicherheit für den pharmazeutischen Hersteller wäre eine konkretere Beschreibung des Vorgehens sicherlich wünschenswert. Grundsätzlich stellt sich jedoch auch hier die Frage nach der ethisch-moralischen Legitimation des IQWiG, solche Werteentscheidungen zu treffen. Bislang wird nämlich in dieser Diskussion außer Acht gelassen, wo die Präferenzen des betroffenen Patienten in einer bestimmten Situation liegen. Im Falle einer schwerwiegenden, zum Tode führenden Krankheit, können in der Wahrnehmung des Patienten Nebenwirkungen durch eine bestimmte Verlängerung des Lebens möglicherweise aufgewogen werden. Solche Präferenzen können von Erkrankung zu Erkrankung je nach Schweregrad der Erkrankung und der möglichen Nebenwirkungen selbst sehr stark variieren, aber auch innerhalb einer Erkrankung können sich die individuellen Wahrnehmungen und Entscheidungen sehr unterscheiden.

Darüber hinaus gibt es für eine Saldierung eines Zusatznutzens mit dem Nebenwirkungsrisiko eines Arzneimittels in der Legaldefinition keine Grundlage:

Nach § 35 a Abs. 1 Satz 2 SGB V ist das Zusatznutzen gegenüber der zweckmäßigen Vergleichstherapie, das Ausmaß des Zusatznutzens und seine therapeutische Bedeutung zu bewerten. Zwar enthält das SGB V keine Legaldefinition des Zusatznutzens, aber § 35 a Abs. 1 Satz 7 Nr. 2 SGB V ermächtigt das BMG, in einer Rechtsverordnung die Grundsätze für die Bestimmung des Zusatznutzens festzulegen. Von dieser Ermächtigung hat das BMG Gebrauch gemacht. § 2 Abs. 3 AMG-NutzenV definiert den Nutzen als für den Patienten relevanten therapeutischen Effekt, insbesondere hinsichtlich der Verbesserung des Gesundheitszustandes, der Verkürzung der Krankheitsdauer, der Verlängerung des Überlebens, der Verringerung von Nebenwirkungen oder einer Verbesserung des Lebensqualität. Der Zusatznutzen wird nach § 2 Abs. 4 AM-NutzenV definiert als ein Nutzen i. S. d. Abs. 3, der quantitativ oder qualitativ höher ist als ein Nutzen, den die zweckmäßige Vergleichstherapie aufweist. Eine Saldierung des festgestellten Zusatznutzens mit dem Nebenwirkungsrisiko eines Arzneimittels findet daher in diesen Legaldefinitionen keine Grundlage. Vielmehr sieht § 5 Abs. 5 Satz 1 AM-NutzenV vor, dass der Zusatznutzen festgestellt wird als Verbesserung der Beeinflussung patientenrelevanter Endpunkte zum Nutzen gem. § 2 Abs. 3. Dass hier Nutzen und Risiken des Arzneimittels im angemessenen Verhältnis zueinander stehen, ergibt sich zwingend aus der arzneimittelrechtlichen Zulassung, weil bei einem ungünstigen Nutzen-Risiko-Verhältnis keine arzneimittelrechtli-che Zulassung erteilt worden wäre (vgl. für die deutsche Rechtslage § 25 Abs. 2 Nr. 5 AMG). Diese Bewertung der Zulassungsbehörde ist auch für die sozialrechtliche Nutzenbewertung von GBA und IQWIG vorgreiflich (vgl. §§ 5 Abs. 3 Satz 2, 7 Abs. 2 Satz 6 AM-NutzenV).

Zu 3.3.3. Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V

Das IQWiG betont, dass die Nutzenbewertung auf Basis der Standards der evidenzbasierten Medizin und der Gesundheitsökonomie erfolgt. (IQWiG 2013) Auch dieser Aussage muss man deutlich widersprechen. Insbesondere die Kategorisierung des Zusatznutzens in Ausmaß und Wahrscheinlichkeit und deren Herleitung folgt weder internationalen Standards, noch wird dies in vergleichbarer Weise in anderen Ländern praktiziert. Auch die Bewertung der Kosten folgt nicht bekannten Standards der Gesundheitsökonomie. Als gesundheitsökonomische Instrumente werden in anderen Ländern beispielsweise Cost-effectiveness-Analysen, Cost-utiliy- Analysen oder Budget Impact Modelle von HTA-Behörden anerkannt. Der deutsche Weg einer stark vereinfachten Kostenminimisierungsanalyse dagegen lässt sich kaum als internationaler Standard bezeichnen, wird von anderen HTA- Behörden nicht in ähnlicher Form angewendet und spiegelt nur einseitig die durch die Erkrankung verursachten Kosten wider.

Zur Feststellung des Ausmaßes des Zusatznutzens verwendet das Institut bei binären Endpunkten vorrangig relative Effektmaße, da diese stabiler seien als absolute wie z.B. Differenzen. (IQWiG 2013) Das vom Institut in den bisherigen Dossierbewertungen präferierte relative Effektmaß ist das relative Risiko, obwohl in der bisherigen Dossiervorlage/ Verfahrensordnung die Odds Ratio und in der neuen Verfahrensordnung beides angeraten wird.

Im Gegensatz zur Odds Ratio hat das relative Risiko den Nachteil, dass die Ergebnisse wesentlich von der Festlegung abhängen, ob man das Ereignis oder das entsprechende Gegenereignis betrachtet. Das Institut empfiehlt, diese Entscheidung auf inhaltlicher Grundlage zu treffen, lässt aber völlig offen, wie dies geschehen soll. Dies öffnet die Tür für

ergebnisgetriebene Festlegungen. Durch die Verwendung der Odds Ratio als Abstandsmaß für binäre Endpunkte hätte dieses Problem vollständig vermieden werden können.

Die Tatsache, dass künftig in Dossiers pharmazeutischer Unternehmer beide Abstandsmaße berichtet werden sollen, beinhaltet, neben zusätzlicher Arbeit und einer verringerten Übersichtlichkeit im Dossier, insbesondere für Meta-Analysen und indirekte Vergleiche das Risiko inkonsistenter Ergebnisse zwischen beiden Abstandsmaßen (z.B. kann in einzelnen Fällten das 95% Konfidenzintervall für das Relative Risiko evtl. die 1 enthalten, nicht aber das 95% Konfidenzintervall für das Odds Ratio).

Im Falle sonstiger stetiger Zielgrößen wäre es an dieser Stelle wünschenswert zu erfahren, wie das IQWiG zur Dichotomisierung am Median der gepoolten Behandlungsgruppen steht, da es das Ausmaß eines Zusatznutzens nur an Hand binärer Endpunkte, die ggf. aus stetigen abgeleitet werden müssen, beurteilen will.

Zu 7.3.8 Meta-Analysen

Zu B) Heterogenität

Bei der Durchführung von Meta-Analysen kann es trotz Einbeziehung vergleichbarer Studien zu heterogenen Ergebnissen kommen, die, wie vom IQWiG gefordert, untersucht werden müssen. Neben dem Testen auf Vorliegen von Heterogenität mit einem Signifikanzniveau von 0,1 bis 0,2, soll auch das Ausmaß der Heterogenität quantifiziert werden. Hier schlägt das IQWiG das I^2-Maß vor und verwendet zur Einschätzung der Heterogenität vorgeschlagene Kategorien aus dem Cochrane-Handbook. (IQWiG 2013) Die aus dem Cochrane-Handbook übernommene und übersetzte Einschätzung der Heterogenität anhand des I^2-Maßes hat den gravierenden Nachteil, dass sich die Bereiche zur Definition der Attribute überlappen. Z.B. I^2 zwischen 0-**40%** wahrscheinlich unbedeutende Heterogenität, I^2 zwischen **30**-**60%** mittelmäßige Heterogenität, I^2 zwischen **50**-**90%** substanzielle Heterogenität und I^2 zwischen **75**-100% erhebliche Heterogenität. Es wäre aus unserer Sicht wünschenswert, eine eindeutige Klassifizierung von Heterogenität zu verwenden.

Zu C) Subgruppenanalysen im Rahmen von Meta-Analysen

Falls es bei Subgruppenanalysen Belege für unterschiedliche Effekte in den Subgruppen (p-Wert des Heterogenitäts- oder Interaktionstest < 0,20) gibt, so sollen gemäß IQWiG die Ergebnisse aller Subgruppen nicht zu einem gemeinsamen Effektschätzer gepoolt werden. Dagegen sollen bei mehr als zwei Subgruppen die paarweise statistischen Tests auf das Vorliegen von Subgruppeneffekten durchgeführt und Paare, die zum Signifikanz-Niveau 0,20 nicht statistisch signifikant sind, zu einer Gruppe zusammengefasst werden. (IQWiG 2013) Bei mehr als zwei Subgruppen ist in paarweisen Heterogenitätstests die Eigenschaft, dass der p-Wert ≥ 0,20 ist und es damit keinen Hinweis auf unterschiedliche Effekte gibt, nicht zwingend transitiv. Beispielsweise kann für die Subgruppenpaarungen A mit B und für B mit C jeweils $p \geq 0{,}20$ erfüllt sein. Damit gibt es keinen Hinweis auf unterschiedliche Effekte von A und B und von B und C. Möglich ist jedoch immer noch der Fall, dass es für die Subgruppenpaarung A mit C einen Hinweis auf unterschiedliche Effekte von A und C gibt. Das Institut beantwortet nicht die Frage, wie in dieser Situation verfahren werden soll.

Darüber hinaus ist auch zu betonen, dass die Festlegung des p-Wertes auf 0,2 als Cut-off für die Bewertung des Vorliegens von Heterogenität in der Literatur durchaus konträr diskutiert wird und auch niedrigere Werte als Cut-off verwendet werden.

Im Zusammenhang mit Subgruppenanalysen versäumt das IQWiG eine adäquate Diskussion des Multiplizitätsproblems, d.h. der Tatsache, dass mit wachsender Anzahl von Heterogenitäts- bzw. Interaktionstests auch das Risiko für falsch positive Hinweise auf Heterogenität anwächst.

Zu D) Geringe Zahl von Ereignissen

Laut IQWiG ist bei der Durchführung von Meta-Analysen von binären Daten das Vorhandensein von sogenannten Nullzellen, also die Beobachtung von keinem einzigen Ereignis in einer Interventionsgruppe einer Studie, ein Problem, für deren Lösung es aber vom IQWiG vorgeschlagene adäquate Methoden gibt. Bei Vorliegen von Doppelnullstudien allerdings (also die Beobachtung von keinem einzigen Ereignis in beiden Interventionsgruppen) ist die Verwendung relativer Effektmaße nicht möglich. Das Institut rät, derartige Studien möglichst nicht aus Meta-Analysen auszuschließen, sondern stattdessen auf absolute Risikodifferenzen auszuweichen. (IQWiG 2013) Die Frage, wie auf Grundlage einer Meta-Analyse von Risikodifferenzen das Ausmaß eines Zusatznutzens bewertet werden kann, lässt das Institut hingegen unbeantwortet.

In diesem Zusammenhang sollte auch Stellung bezogen werden, wie das Institut zu NNT als Abstandsmaß steht.

Zu Anhang: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens

Eine Anpassung der Kriterienliste gemäß AM-NutzenV ist nachvollziehbar. So spielt der Begriff „Heilung“ in erster Linie bei Antibiotika-Studien eine Rolle, die aus ethischen Gründen häufig als Nicht-Unterlegenheitsstudien geplant werden – damit ist im direkten Vergleich ein Zusatznutzen i.d.R. nicht erreichbar, obwohl für neue Antibiotika mit neuartigen Angriffspunkten ein allgemein anerkannter Bedarf besteht.

Auch eine Ergänzung der Zielgrößenkategorisierung um gesundheitsbezogene Lebensqualität ist zu begrüßen. Zusätzlich aber wäre es auch wünschenswert, bestimmte Surrogatparameter, die auch von nationalen und internationalen Zulassungsbehörden akzeptiert werden, als patientenrelevant in den Zielgrößenkatalog mit aufzunehmen. Hinzu kommt, dass die Schweregrade von Symptomen und Nebenwirkungen (jew. schwerwiegend/ schwer oder nicht schwerwiegend/ nicht schwer) nicht operationalisiert sind, und auch an dieser Stelle wieder eine Werteentscheidung durch das IQWiG getroffen wird. Den Zielgrößenkategorien werden die Ausmaßkategorien „erheblich“, „beträchtlich“ und „gering“ zugeordnet und zwar auf Basis von Obergrenzen von 95%- Konfidenzintervallen im Sinne verschobener Hypothesengrenzen. Das IQWiG ist der Ansicht, dass die Forderung der AM-NutzenV nach Berücksichtigung der Krankheitsschwere damit ausreichend umgesetzt sei. (IQWiG 2013) Aufgrund der oben dargelegten fehlenden Operationalisierung schwerwiegender/ nicht schwerwiegender Symptome und Nebenwirkungen ist dies fraglich.

Ein gravierender Nachteil des Konzeptes verschobener Hypothesengrenzen besteht zusätzlich darin, dass das Ausmaß des Zusatznutzens neben dem Punktschätzer erheblich von der Fallzahl der Studien abhängt – ein geringerer Abstand zur zweckmäßigen Vergleichstherapie kann durch eine höhere Fallzahl zumindest teilweise kompensiert werden, da die Länge des Konfidenzintervalls mit steigender Fallzahl abnimmt, so dass im Endeffekt das gleiche Ausmaß des Zusatznutzens attestiert wird. Dieses Vorgehen impliziert also eine Vermengung des Therapieeffektes mit der Präzision seiner Schätzung, die zu einem medizinisch schwierig zu bewertenden und zwischen Dossiers zu vergleichenden Urteil über das Ausmaß des Zusatznutzens führt.

Das IQWiG legt dann eine Rasterung der Schwellenwerte von 0,05, d.h. auf Zwanzigstel, fest. Eine Rationale dafür wird nicht gegeben, sondern es wird nur festgestellt, dass dies in der Praxis anwendbar sei. (IQWiG 2013)

Es bleibt weiterhin unklar, warum eine Rasterung der tatsächlichen (gewünschten) Effekte für das relative Risiko im Abstand von Sechsteln, d.h. in Form einer arithmetischen Folge (konstante Abstände) durchgeführt wurde. Nachdem die Punktschätzer für relatives Risiko und Hazard Ratio logarithmisch normalverteilt sind, erscheint eine Rasterung der gewünschten Effekte gemäß einer geometrischen Folge, d.h. mit konstantem Quotienten benachbarter Effekte, viel naheliegender und hätte insbesondere die extreme Anforderung für schwerwiegende Symptome an einen erheblichen Zusatznutzen abgemildert (gewünschtes RR=0,31 statt 0,17).

Der einzige Ankerpunkt für diese Rasterung, der der Arbeit von Djulbegovic et al. entnommen wurde, weist zudem äußerst schwache Evidenz auf: Diese Literaturstelle ist vielmehr eine Untersuchung onkologischer Therapiestudien und nennt die Halbierung des Mortalitätsrisikos ohne Begründung und lediglich als Teil einer Arbeitsdefinition für Durchbruchsinnovationen, die sie selber als arbiträr kennzeichnet: „We determined the proportion of discoveries that were ‚breakthrough interventions'. This was arbitrarily defined as interventions judged by the original researchers to be so beneficial that they should immediately become the new standard of care or that had an effect size so large that they reduced the death rate by 50% or more (ie, the HR for death was 0.5 or less)" (Djulbegovic et al. 2008)

In diesem Zusammenhang bleibt festzuhalten, dass in Djulbegovic et al. 116 Studien identifiziert werden, in denen die Autoren von einem innovativen Durchbruch berichten, aber lediglich 12 Studien der Subdefinition der Halbierung des Mortalitätsrisikos entsprechen:

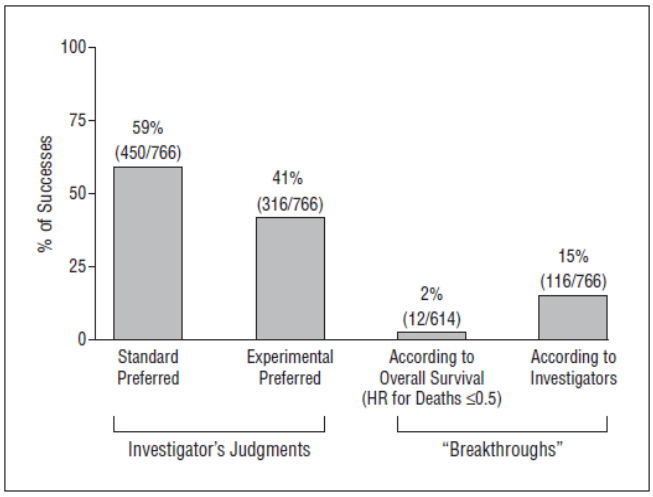


Figure 3. Distribution of outcomes according to the published judgment of the original researchers. "Breakthrough" treatments were defined as those that should replace the existing standard of care or that reduced the death rate by 50% or more. HR indicates hazard ratio.

Übernommen aus: Djulbegovic et al. 2008

Zu Anhang: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens- Detaillierte methodische Rationale für die Festlegung der Schwellenwerte

An dieser Stelle beschreibt das IQWiG das methodische Vorgehen zur Festlegung der unterschiedlichen Schwellenwerte auf Basis der verschobenen Hypothesengrenzen. Das IQWiG geht dabei vom Regelfall zweier vorliegender pivotaler Studien aus. (IQWiG 2013) In onkologischen Indikationen kann nicht vom Regelfall des Vorliegens von zwei pivotalen Studien ausgegangen werden, da eine Reproduktion einer erfolgreichen Studie ethisch nicht vertretbar wäre. Das Institut führt hier nicht aus, wie in dieser Situation ein Beleg für einen Zusatznutzen erreicht werden kann.

Der im Anhang A zur Ticagrelor Nutzenbewertung verwendete Algorithmus zur Herstellung eines Zusammenhangs zwischen tatsächlichem Effekt und Schwellenwert ist nach Aussage des IQWiG nicht publiziert, weshalb die tatsächlichen Effekte per Monte-Carlo-Simulation nachgeprüft wurden. (IQWiG 2013)

Das Institut legt auf Grund der durchgeführten Monte-Carlo-Simulationen dar, dass es für schwerwiegende Symptome und gleichwertige Zielgrößen de facto einer Risikoreduktion auf etwa ein Viertel bis ein Drittel bedarf, um einen erheblichen Zusatznutzen nachzuweisen. (IQWiG 2013) Dieses Ergebnis steht in deutlichem Gegensatz zu der Annahme RR=0,17, die der theoretischen Herleitung für den Schwellenwert in dieser Situation zu Grunde liegt und belegt an diesem Beispiel, dass die von Herrn Prof. Röhmel formulierte Kritik an der Herleitung der Schwellenwerte berechtigt ist. So konnte Prof. Röhmel die angegebenen Schwellenwerte aus den angegebenen wahren anzunehmenden Risikoverhältnissen und bei Verdopplung der Fallzahl rechnerisch nicht nachvollziehen, sondern kam zu anderen Ergebnissen - (s. Anhang Diskussionspapier zum Vorschlag des IQWiG zur Bewertung des Ausmaßes des Zusatznutzens im Rahmen der Nutzenbewertung von Arzneimitteln nach §35a SGB V; 21. Mai 2012)

Literatur:

- Djulbegovic B et al. Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006. Arch Intern Med 2008; 168(6): 632-642.
- EMEA. The European Agency for the Evaluation of Medicinal Products. Points to consider on application with 1. Meta-analyses; 2. one pivotal study (CPMP/EWP/2330/99); 2001
- Gemeinsamer Bundesausschuss. Mündliche Anhörung gemäß 5. Kapitel, § 19 Abs. 2 Verfahrensordnung des G-BA; hier: Wirkstoff Ticagrelor. Stenographisches Wortprotokoll; Berlin, 17.11.2011
- Grade Working Group. Education and Debate. Grading quality of evidence and strength of recommendations. BMJ 2004; 328: 1490
- IQWiG. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1. Köln, 18.04.2013
- IQWiG. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Link: https://www.iqwig.de/de/presse/pressemitteilungen/pressemitteilungen/ueberarbeitung_methodenpapier_iqwig_stellt_aenderungen_am_methodenpapier_zur_diskussion.3633.html; entnommen am 07.05.2013 (2013a)
- Röhmel J. Diskussionspapier zum Vorschlag des IQWiG zur Bewertung des Ausmaßes des Zusatznutzens im Rahmen der Nutzenbewertung von Arzneimitteln nach §35a SGB V"; 21. Mai 2012

A.1.23 – Verband Forschender Arzneimittelhersteller e. V. (vfa)

Stellungnahme des vfa zum Entwurf des IQWiG „Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“ vom 18.04.2013

22.05.2013

Dr. Ch.-Markos Dintsios

vfa
Senior Referent HTA & Gesundheitsökonomie
Hausvogteiplatz 13, D-10117 Berlin
[redacted]

Gliederung:

Stellungnahme des vfa zum Entwurf des IQWiG:
„Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“ vom 18.04.2013

Die Stellungnahme des Verbandes forschender Arzneimittelhersteller e.V. (vfa) bezieht sich sowohl auf methodische als auch auf formale Aspekte sowie einzelne Punkte zur Umsetzung der im Entwurf vom IQWiG zur Stellungnahme freigegebenen „Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“ vom 18.04.2013. Die Stellungnahme erfolgt chronologisch entlang des vom IQWiG vorgelegten Entwurfs.

A) Entwurf zu den Abschnitten 2.1.1 Bericht und 2.2.3 Review der Produkte des Instituts

Bei Durchsicht der beiden Abschnitte 2.1.1 Bericht und 2.2.3 Review der Produkte des Instituts fällt als Änderung im Entwurf des IQWiG auf, dass nun der Passus zum externen Review des Vorberichts in Abschnitt 2.2.1 *„Zudem wird als weiterer Schritt der Qualitätssicherung der Vorbericht einem oder mehreren externen Reviewern (siehe Abschnitt 2.2.3) mit ausgewiesener methodischer und / oder fachlicher Kompetenz vorgelegt“* (S. 15 Ver. 4.0) gestrichen wurde und im Abschnitt 2.2.3 der Passus *„Darüber hinaus wird im Verlauf der Erstellung von Berichten und z. T. auch von Gesundheitsinformationen ein externes Reviewverfahren als weiterer Schritt der Qualitätssicherung durchgeführt*“ (S. 28 Ver. 4.0) dahingehend abgewandelt wurde, dass das externe Review des Vorberichts nun nur optional umgesetzt wird: *„Darüber hinaus kann im Verlauf der Erstellung von Berichten und z. T. auch von Gesundheitsinformationen (siehe 2.1.5) ein externes Reviewverfahren als optionaler weiterer Schritt der Qualitätssicherung durchgeführt werden“* (Entwurf Ver. 4.1). Der vfa ist der Auffassung, dass ein externes Review durchaus der Qualitätssicherung der IQWiG Produkte dient und dass dieser Schritt bei allen IQWiG Vorberichten beibehalten werden sollte. Dies v. a. in Anbetracht der Tatsache, dass wie auch vom IQWiG angemerkt, die Auswahl externer Reviewer primär auf Basis ihrer methodischen und/oder fachlichen Expertise bis dato erfolgte und somit eine zweite Sicht auch aus fachlich klinischer Perspektive auf die entsprechenden IQWiG Produkte ermöglichte. Da sich die klinischen Fachdisziplinen im Rahmen der schon seit Jahren nun fortschreitenden Ausdifferenzierung immer mehr spezialisieren und nicht davon auszugehen ist, dass dieses klinische Fachwissen bei den IQWiG Mitarbeitern immer zur Verfügung steht bzw. stehen wird – was weder von einem Institut wie dem IQWiG verlangt werden kann noch realisierbar wäre – sollte das

externe Review nicht nur optional erfolgen. Ferner ist nicht nur medizinischen Fachwissen erforderlich sondern auch und vor allem Kenntnisse über den Versorgungsalltag hinsichtlich der Anwendung bzw. Umsetzung der zu bewertenden Interventionen im realen Versorgungskontext. Insbesondere auch dieser Punkt dürfte ein externes Review erforderlich machen. In der Medizin, die wissenschaftstheoretisch immer noch keine exakte Wissenschaft darstellt („ärztliche Kunst“), ist oft davon auszugehen, dass sich mehrere Schulen mit unterschiedlichen klinischen Herangehensweisen und Interpretationen auch auf nationaler Ebene herausbilden. Der Disput über die klinische Wertigkeit der Albuminurie in der deutschen Diabetologie soll hier nur als ein Beispiel für das oben beschriebene Phänomen dienen (Schmieder 2010, Richter 2002). Somit erhöht ein externes Review die Chancen auf eine ausgewogene klinisch fachliche Darstellung der Bewertungen. Das externe Review in seiner obligatorischen Umsetzung erscheint einer optionalen Umsetzung somit überlegen. Um allerdings auch ein selektives Review bzw. eine interessengeleitete Auswahl der Reviewer zu vermeiden, wäre – dem Gedanken des Transparenzgebotes Folge leistend – eine klare Institutionalisierung des Reviews bezüglich der Auswahlkriterien der geladenen Reviewer, der einzuhaltenden Fristen, der Offenlegung der Beratungsfragen und der -ergebnisse wünschenswert.

B) 3.1.4 Endpunktbezogene Bewertung (neuer Abschnitt)

In seinem neuen Abschnitt 3.1.4 Endpunktbezogene Bewertung führt das IQWiG u.a. die qualitative Ergebnissicherheit und das Prädiktionsintervall zur Darstellung der Heterogenität in sein Methodenpapier als Kriterien bzw. Instrumente einer endpunktbezogenen Bewertung ein. Hinsichtlich der geringen qualitativen Ergebnissicherheit sei auf diesem Wege besonders betont, dass es Konstellationen gibt, die aufgrund ethischer, methodischer und patientenpräferenzbedingter Gründe keine Randomisierung ermöglichen und somit durchaus angebracht sein können. In diesen Fällen die Ergebnissicherheit pauschal und regelhaft als qualitativ gering einzuschätzen, ohne die Machbarkeit einer Randomisierung zu überprüfen, erscheint nach Auffassung des vfa ungerechtfertigt. Hinzu können bezogen auf eine langfristige Betrachtung prohibitiv hohe Kosten für die Durchführung einer RCT kommen. So könnten Studien in solchen Fällen mit einer geringeren qualitativen Ergebnissicherheit gekennzeichnet werden, in welchen eine Randomisierung auch möglich

gewesen wäre. Das IQWiG beginnt seine Ausführungen mit folgender Aussage: *„die Nutzenbewertung und die Einschätzung der Stärke der Ergebnis(un)sicherheit orientieren sich an internationalen Standards der evidenzbasierten Medizin, wie sie z. B. von der GRADE-Gruppe erarbeitet werden“* und verweist auf entsprechende Literatur. Dabei bezieht sich das IQWiG in seinen weiteren Ausführungen auf die Abstufungen zur Belegbarkeit des (Zusatz-)Nutzens aufgrund der Beleglage der Ergebnissicherheit und weist sowohl auf die qualitative als auch auf die quantitative Ergebnisunsicherheit hin. Interessanterweise wird zwar vom IQWiG auf Kriterien der GRADE-Gruppe verwiesen, die zu einer abwertenden Einschätzung der Ergebnissicherheit führen können, nicht aber auf Kriterien derselben Gruppe, die für Sonderfälle auch eine Aufwertung der Ergebnissicherheit vorsehen bzw. führen können (Atkins et al. 2004). Der Verweis auf internationale Standards der evidenzbasierten Medizin sollte nach Auffassung des vfa nicht selektiv erfolgen und bei Nennung von international etablierten Initiativen auch vollständig auf deren vorgeschlagenes Vorgehen eingegangen werden. Im Sinne einer umfassenden Herleitung, sollten daher vom IQWIG hier auch die von der GRADE-Gruppe als Sonderfälle vorgesehenen Kriterien diskutiert werden. Ferner sei hier hervorgehoben, dass eine endpunktbezogene Bewertung zumindest für Arzneimittel bereits im Rahmen ihrer Zulassung erfolgt ist. Damit stellt sich die Frage, inwieweit bei Befolgung der internationalen Standards der evidenzbasierten Medizin, denen sich auch die Zulassungsbehörden auf nationaler und internationaler Ebene verpflichtet fühlen, zu andersgearteten Ergebnissen auf Basis der selben Studienlage führen sollte. Dies ist für den vfa und seine Verbandsmitglieder insoweit von Interesse, als im Rahmen der frühen Nutzenbewertung von Arzneimitteln auf Basis der vorhandenen Zulassungsstudien einerseits die Bindungswirkung der Zulassung durch das Vorgehen des IQWiG konterkariert werden kann, andererseits bei der Umsetzung der internationalen Standards der evidenzbasierten Medizin es nicht zu unterschiedlichen (endpunktbezogenen) Bewertungen zumindest hinsichtlich der Ergebnissicherheit kommen sollte. Dies führt auch die Logik jeglicher Standards ad absurdum, wenn diese kontextbezogen unterschiedlich interpretiert und implementiert würden.

In diesem Abschnitt verweist das IQWiG auch in Abhängigkeit von der Fragestellung hinsichtlich des Vorhandenseins eines (Zusatz-)Nutzens auf gut begründete Definitionen von **Irrelevanzbereichen**. Zwar ist der entsprechende Abschnitt (7.3.6) nicht Gegenstand des vorliegenden Entwurfs zur Aktualisierung der Allgemeinen Methoden, es sollte nach

Auffassung des vfa dennoch darauf hingewiesen werden, dass dies methodisch durchaus ein umstrittenes Feld darstellt, weil es sich primär um arbiträre Setzungen handelt, also Irrelevanzbereiche als Behelfsmittel eingesetzt werden und diesen somit nicht eine überragende Bedeutung beigemessen werden sollte. Das IQWiG selber wagt – worauf im Weiteren noch eingegangen wird – in seinem Vorgehen zur Klassifizierung des Zusatznutzens Setzungen von „Irrelevanzbereichen" über Konfidenzintervallgrenzen bzw. -schwellen, die weder national oder international diskutiert, noch als gut begründet angesehen werden können.

Bei der Bestimmung des Vorliegens gleichgerichteter Effekte führt das IQWiG das **Prädiktionsintervall** in seinen Entwurf zur Aktualisierung der Methoden ein und stellt einen umfangreichen Regelkatalog für unterschiedliche Situationen auf. In diesen Regelkatalog gehen das Gesamtgewicht der Studien, die statistische Signifikanz der Ergebnisse sowie das Gewicht der auf statistisch signifikanten Ergebnissen fußenden Studien ein. Wir bitten das IQWiG, die Rationale für die Einführung des auf Basis von Meta-Analysen mit zufälligen Effekten hergeleiteten Kriteriums Prädiktionsintervall für die Beurteilung bzw. Einstufung der Gleichgerichtetheit in seinem Methodenentwurf zu ergänzen. Zwar gibt es in der Literatur theoretische Abhandlungen hierzu, diese können jedoch nicht als etablierte Standards angesehen werden. Dafür spricht u. a. auch die Tatsache, dass Prädiktionsintervalle in den Standard-Softwarepaketen bisher nicht implementiert sind. Auch andere HTA Agenturen verwenden das Prädiktionsintervall nicht in dieser Weise, was ebenfalls als Indiz für seine Nicht-Anwendung angesehen werden kann. Eine notwendige Rationale zur Vermeidung des Eindrucks willkürlicher Festlegungen fehlt auch zu den entsprechenden Setzungen hinsichtlich des Gesamtgewichts der **„gerichteten" Studien** ($\geq$ 80%), der Anzahl mit signifikanten Ergebnissen (N $\geq$ 2) und des Mindestgewichts (50%) der auf statistisch signifikanten Ergebnissen basierenden Studien. Der vfa ist der Auffassung, dass das vom IQWiG in seinem Methodenentwurf neu eingeführte Prädiktionsintervall und die Schwellenwerte zum Gesamtgewicht gerichteter Studien mit einer Aussage hinsichtlich ihrer Applikation bzw. Implementierung auf internationaler Basis belegt werden sollten, denn nicht alles was methodisch entwickelt wurde, stellt auch automatisch einen internationalen Standard dar. Für die entsprechenden Schwellenwerte, die als Bewertungs- bzw. Einstufungskriterien eingesetzt werden, sollte auch die dahinterstehende Rationale mitgeliefert werden. Ferner sollten diesbezüglich auch Stellungnahmen nationaler (z. B.

GMDS) und internationaler methodisch-statistischer Akteure in die Betrachtung mit einbezogen werden, um die methodischen (Weiter-)Entwicklungen auch unter IQWiG-externen Experten kritisch zu diskutieren. Bei der Entwicklung von Methoden sollte auch im Augenmerk bleiben, dass diese umsetzbar, zielführend und sinnvoll sind, da eine ausschließlich akademisch-theoretische Verortung der neu im Entwurf eingebrachten Vorgehensweisen auch ihre Machbarkeit und Transparenz mit berücksichtigen und gewährleisten sollte. Der vfa kann sich aufgrund des fehlenden Gleichgewichtes zwischen potentiellem Informationsgewinn und methodischen Schwächen beim vorgeschlagenen Vorgehen des IQWiG nicht des Eindrucks entledigen, dass hier neue Hürden geschaffen werden, die ohne adäquate wissenschaftliche Untermauerung eine regelhafte Herabstufung der Ergebnissicherheit erleichtern und somit zumindest für die Empfehlungen des IQWiG im Rahmen der (frühen) Nutzenbewertungen maßgeblichen Einfluss nehmen können. So können Konstellationen bei der Zuordnung von Aussagen aus mehreren Studien hinsichtlich der Ergebnissicherheit unter Zuhilfenahme der Richtung der Effekte zu einem Resultat führen, dass selbst bei mäßiger qualitativer Ergebnissicherheit bei gleichzeitig mäßiger Gleichgerichtetheit nur ein Anhaltspunkt zugesprochen wird, was das IQWiG Vorgehen nach Auffassung des vfa äußerst konservativ gestaltet und die Gefahr birgt, dass der Anhaltspunkt unnötigerweise anstelle eines Hinweises sehr oft vergeben werden kann.

Weiter gilt es, zum neuen Abschnitt anzumerken, dass in der Tabelle 2 „Anforderungen für die Beleglage“ ein Verweis auf homogene bzw. heterogene Effekte aus Meta-Analysen erfolgt, ohne auf entsprechende Kriterien (beispielsweise I^2-Wert) und deren Grenzwerte hier explizit einzugehen sowie die Konstellation mit **einer Studie** für einen Beleg nicht aufgenommen wurde, die im Text allerdings wie folgt beschrieben wird: „*Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und ihre Ergebnisse besondere Anforderungen zu stellen*“ und damit auch gegeben ist. Hier wäre anzumerken, dass die Bindungswirkung der Zulassung bei Vorliegen einer Studie nicht ignoriert werden sollte. Der Verweis auf die „Points to consider“ des EMA Vorgehens (EMA 2001) bei Vorliegen einer pivotalen Studie reicht nicht unbedingt alleine aus, da selbst bei von der EMA akzeptierten Bedingungen hierfür, im Rahmen der frühen Nutzenbewertung beispielsweise das IQWiG die EMA Kriterien in Abrede stellt bzw. eine eigenwillige Interpretation dieser, oft auch mit Verweis auf das Sozialgesetzbuch wagt (siehe dazu beispielsweise Ausführungen von Frau Dr. Wieseler in der mündlichen Abhörung zur

frühen Nutzenbewertung von Abirateronacetat nach §35 a SGB V – festgehalten im entsprechenden Wortprotokoll vom 07.02.2012 – wonach wortwörtlich: „Die Ergebnissicherheit einer Studie und der Aussagen aus einer Studie hängt eben nicht von der relativen Patientenzahl ab, also von der Patientenzahl im Vergleich zur verfügbaren Population, sondern eher von der absoluten Patientenzahl. Also angenommen, Sie hätten eine sehr kleine Patientenzahl und hätten diese kleine Patientenzahl zu einem großen Anteil in Ihre Studie eingeschlossen, dann kämen Sie damit trotzdem nicht zu einer ergebnissicheren Aussage“, was zu Ende gedacht hieße, dass selbst bei einem Einschluss von über 90% der Grundgesamtheit in einer Studie keine Ergebnissicherheit für einen Beleg gewährleistet wäre, wenn diese Grundgesamtheit klein ausfällt). Wünschenswert wäre nach Auffassung des vfa eine klare Positionierung des IQWiG, ob die entsprechenden „Points to consider“ der EMA nun erschöpfend sind oder nicht und falls nicht, welche weiteren Kriterien erfüllt sein sollten. Weitere Kriterien sind nach Auffassung des vfa zu ergänzen. Hierzu zählen beispielsweise dramatische Effekte aus einer randomisierten Studie hoher Ergebnisqualität im direkten Vergleich mit der zweckmäßigen Vergleichstherapie. Ein endpunktbezogener Beleg ist auch dann zu gewähren, wenn die Zulassungsbehörde basierend auf einer randomisierten Studie hoher Ergebnisqualität im direkten Vergleich mit der zweckmäßigen Vergleichstherapie unter "4.1 Anwendungsgebiete" endpunktbezogene patientenrelevante Aussagen aufnimmt, beispielsweise "zur Vermeidung von Schlaganfällen". Nur ein "Hinweis" auf einen Zusatznutzen für diesen Endpunkt würde diese regulatorische Maßnahme konterkarieren.

Des Weiteren wird in der Tabelle 2 hinsichtlich der Gleichgerichtetheit der Effekte auf $\geq$ 2 Studien hingewiesen, was auf > 2 geändert werden müsste, da bei ausschließlich 2 gleichgerichteten Studien die Gleichgerichtetheit nicht Eingestuft werden kann, da sie eindeutig ausfällt.

Hinsichtlich der vom IQWiG geäußerten weiteren Faktoren, die seine Einschätzung beeinflussen können, sei explizit darauf hingewiesen, dass dieser Passus eher abstrakt gehalten wird und dass bei Zweifeln an der Übertragbarkeit auf die Behandlungssituation in Deutschland nach Auffassung des vfa doch eher auf Westeuropa auch aus rein methodischen Gründen anstelle von Deutschland verwiesen werden sollte, um eine Fehlinterpretation aufgrund multipler Effekte zu vermeiden. Zudem wäre es nach Auffassung des vfa notwendig, durch eindeutige Kriterien die Einschätzung einer Übertragbarkeit zu

operationalisieren. Das regionale Poolen erscheint hier eher als gangbarer Weg, um statistische Fehlschlüsse – bedingt beispielsweise durch die die hinlänglich bekannten Gefahren des multiplen Testens, die auf nationalen Auswertungen bzw. sogar auf Auswertungen auf Zentrumsebene beruhen, zu verringern.

C) 3.1.5 Zusammenfassende Bewertung

Der Passus „*Die gleichzeitige Würdigung von Nutzen und Schaden wird themenspezifisch konkretisiert und sollte – wenn dies prospektiv möglich ist – im Berichtsplan oder andernfalls im Vorbericht beschrieben werden. Eine quantitative Gewichtung unter Verwendung von Summenscores oder Indizes sollte prospektiv zum Zeitpunkt der Auswahl der zu untersuchenden Endpunkte erfolgen*“ klingt widersprüchlich, betrachtet man erstens das Vorgehen des IQWiG im Rahmen der frühen Nutzenbewertung, wo kein Raum für eine quantitative Gewichtung hinsichtlich der Nutzen- und Schadenaspekte erfolgt und zweitens davon ausgegangen werden kann, dass Präferenzstudien auch bei Bewertungen außerhalb des §35a SGB V meistens zum Zeitpunkt der Auswahl der untersuchten Endpunkte nicht fertig sein werden. Ferner erfolgen Gewichtungen mit Koeffizienten oder unter Anwendung der marginalen Grenzrate der Substitution und nicht anhand von Summenscores. Somit haben wir es in Abhängigkeit von der Zielsetzung bzw. dem Auftrag und dem entsprechenden Verfahren, das hierfür angewendet wird, mit zwei unterschiedlichen Einbringungen von potenziellen Gewichtungen für Nutzen- und Schadensaspekte zu tun, die dieses Vorgehen rein methodisch betrachtet nicht rechtfertigen, da aus der methodischen Beurteilung betrachtet, es keinen Unterschied macht, ob die Nutzenbewertung im Kontext der frühen Nutzenbewertung oder der Kosten-Nutzen Bewertung erfolgt. Eine entsprechende Berücksichtigung quantitativer Gewichtungen sollte auch im Rahmen der frühen Nutzenbewertung erwogen werden.

In Abschnitt 3.1.5 „Zusammenfassende Bewertung“ werden wie in der Version 4.0 (Abschnitt 3.1.4) zwei zum QALY-Ansatz alternative Methoden vorgeschlagen. Dort heißt es: „*Aufgrund der ethischen und methodischen Probleme gerade der häufig verwendeten QALYs sollten alternative Verfahren der multikriteriellen Entscheidungsfindung oder der Präferenzerhebung angewendet werden. Dazu zählen u. a. der Analytic Hierarchy Process (AHP) und die Conjoint-Analyse (CA)*.“

Zuerst sei hier darauf hingewiesen, dass diese Feststellung nicht zielführend ist. Es sind nicht die Präferenzmessmethoden, welche in Konkurrenz zum QALY stehen. Auch die Lebensqualität im Rahmen des QALY kann beispielsweise mittels einer Conjoint Analyse erhoben werden. Es ist die endpunktbasierte Bewertung, die entgegen der Bewertung durch Lebenszeit und Lebensqualität, den Unterschied ausmacht.

Grundsätzlich zu begrüßen ist die ausdrückliche Erwähnung der beiden Verfahren Analytical Hierarchy Process (AHP) und Discrete Choice Analysis oder Conjoint Analysis (CA), wobei es sich hier um eine spezifische Form der CA, der wahlbasierten CA (Discret Choice Analyse) handeln sollte. Beide Verfahren werden jedoch lediglich nebeneinander erwähnt, eine Einordnung in Bezug auf die zusammenfassende Nutzenbewertung sowie eine Graduierung der Verfahren werden nicht vorgenommen. Die beiden Verfahren unterscheiden sich erheblich in ihrer Methodik (Dintsios 2012).

Beim AHP handelt es sich um ein komplexes Verfahren, in dem in mehreren Schritten sowohl qualitative (Präferenzen) wie auch quantitative Aussagen letztlich in einer einzigen Zahl pro Entscheidungsmöglichkeit mittels Matrizenrechnung zusammengefasst werden. Auf Basis eines Rankings wird dann die Hierarchie verschiedener Endpunkte bestimmt werden. Die Anwendung für medizinische Fragestellungen befindet sich teilweise noch im Forschungs- und Entwicklungsstadium, es liegen relativ wenige Publikationen im entsprechenden Kontext vor (z.B. Libartore & Nydick 2008).

Im Einzelnen ergeben sich u.a. folgende kritische Aspekte bei der Anwendung des AHP: (i) Die Metrik beim AHP – Outcome Ordinal oder Kardinal skaliert – ist ungeklärt, was von erheblicher Tragweite sein kann. (ii) Anzahl und Art der einzubeziehenden Endpunkte ist ungeklärt. Diese hängen von der Indikation, den verfügbaren Therapien sowie den verfügbaren Daten ab und nehmen erheblichen Einfluss auf das Ergebnis. (iii) Die einzubeziehenden Perspektiven (z.B. Patient, Arzt) sind nicht im Methodenentwurf definiert und können große Unterschiede aufweisen (Danner et al. 2011, Hummel et al. 2011), was auch von Vorteil sein kann, wenn es um die Berücksichtigung der Patientenperspektive anhand ihrer eigenen Präferenzen geht. (iv) Die jeweiligen Charakteristika der befragten Personengruppe können sich deutlich unterscheiden und somit grundsätzlichen Einfluss auf Präferenzen nehmen, woraus wiederum widersprüchliche Aussagen und Ergebnisse resultieren können. So müsste z.B. geklärt werden, ob bereits erkrankte Patienten oder nicht

erkrankte Probanden befragt werden –wobei Patientenpräferenzen nur bei ersteren erhoben werden können – und in welcher Weise die Befragung durchgeführt werden sollte. Auch die Repräsentativität der befragten Personen spielt hier eine große Rolle für die Fallzahlplanung. Aus Untersuchungen zur Lebensqualität ist bereits bekannt, dass eine Vielzahl von Faktoren, wie z.B. der kulturelle Hintergrund, das Alter oder der Schweregrad einer bereits vorliegenden Erkrankung als Einflussfaktoren fungieren. (v) Schließlich ist der Umgang mit fehlenden Werten („missing values“) noch nicht eindeutig geklärt (Saaty 2008, Ishizaka & Labib 2011). Allerdings kann die Problematik heterogener Präferenzen wiederum bei der CA durch Mixed Logit mit Random Effects statistisch geprüft und berücksichtigt werden.

Da beide Studienformen erst seit wenigen Jahren zu medizinischen Fragestellungen Anwendung finden – wobei sich eine unterschiedliche Literaturbasis mit umfangreicheren Veröffentlichungen zur CA zeigt – und sich daher teilweise noch im Forschungsstadium befinden, können sie noch nicht regelhaft angewandt werden und ihre Aussagen sind entsprechend zu interpretieren.

Angesichts der noch ungeklärten methodischen Fragen halten wir es aber für wichtig, dass diese Verfahren explizit weiterhin Gegenstand der methodischen und indikationsspezifischen Forschung bleiben. Im Rahmen der vergleichenden Bewertung sollten sie optional als zusätzliche Mittel angewendet werden, um eine auf Studienergebnissen basierende Aussage zu unterstützen. Bei ihrer Verwendung in dem Entwurf zum Methodenpapier sollte auf diese Punkte hingewiesen werden. Auch unter Verwendung dieser Methoden wird es kein eindeutiges Ergebnis für alle möglichen Konstellationen geben, aber der Weg zu einer zusammenfassenden Bewertung für alle Beteiligten transparenter und damit nachvollziehbarer.

Zusammenfassend ist der vfa der Auffassung, dass bei dem derzeitigen methodischen Stand, die genannten Verfahren (AHP, CA) eher als zusätzliche optionale Methoden eingesetzt werden sollten, um Bewertungen, die auf mehreren Kriterien beruhen, wozu eindeutig auch die Abwägung von positiven und negativen Effekten im Rahmen der frühen Nutzenbewertung nach §35a SGB V gehört, zu unterstützen. Sie bedürfen der methodischen Weiterentwicklung, was Gegenstand der auch vom IQWiG mitgetragenen weiteren Forschung sein sollte.

D) 3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V und Anhang: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens

Bevor auf die einzelnen, kritischen Punkte zu diesem Abschnitt eingegangen wird, möchte der vfa darauf hinweisen, dass der Versuch des IQWiG im Rahmen der Nutzenbewertung von Arzneimitteln nach § 35a SGB V einen Zusatznutzenklassifizierungsmodus zu entwickeln sicherlich eine methodische und intellektuelle Herausforderung darstellt und das IQWiG hierfür auch gebührenden Respekt verdient. Dennoch ist das vorgeschlagene Vorgehen hinsichtlich seiner methodischen Fundierung und Praktikabilität sowie seiner Legitimierung zu hinterfragen und hier bereits vorwegnehmend aus Sicht des vfa aus mehreren Gründen in seiner jetzigen Form zumindest abzulehnen. Der vfa möchte sich an dem konstruktiven Dialog zur Weiterentwicklung der vorgeschlagenen Methodik aktiv beteiligen und wird weiterhin auch hierzu vom G-BA und dem IQWiG unterbreitete Angebote wahrnehmen. Zuvorderst soll angemerkt werden, dass die Prüfung, ob die vom pharmazeutischen Unternehmer gewählte Vergleichstherapie als zweckmäßig im Sinne des § 35a SGB V und der AM-NutzenV gelten kann, nicht durch das IQWiG erfolgen sollte, da dies eine Festlegung des G-BA ist, deren Herleitung auch vom Entscheidungsträger selber dokumentiert wird und deren Umsetzung und Prüfung aus Legitimationsgründen nicht einem Auftragsinstitut überlassen werden sollte. Es wäre zumindest zu begrüßen, wenn das IQWiG bei von ihm festgestellten Abweichungen auf den G-BA bzw. Hersteller zurückgeht und diese im Einvernehmen klärt. Ferner wird in diesem Abschnitt hinsichtlich der Nutzenbewertung und der Bewertung der Kosten auf die **Standards der evidenzbasierten Medizin** und der Gesundheitsökonomie verwiesen („*Die Nutzenbewertung erfolgt auf Basis der im vorliegenden Methodenpapier beschriebenen Standards der evidenzbasierten Medizin, die Bewertung der Kosten auf Basis der Standards der Gesundheitsökonomie*“), obgleich viele methodische Aspekte der Bewertung international anders gehandhabt werden, so dass dieser Verweis gegenstandslos wirkt. So erfolgt beispielsweise bei der Kostenbewertung den internationalen Standards der Gesundheitsökonomie folgend immer eine Berücksichtigung von cost-offsets, die allerdings vom IQWiG wiederum nicht in seine Bewertung einbezogen werden. Auch werden bei anderen HTA-Agenturen Interaktionstests mit einem Hinweis auf einem Effektmodifikator bei $p = 0,2$ in ihrer Operationalisierung nicht als relevant gedeutet, was auch nicht unbedingt zu den Standards der evidenzbasierten Medizin zuzurechnen ist.

Nach Auffassung des vfa sollte im Entwurf zu den IQWiG Methoden klar beschrieben werden, welche Inhalte von den „proklamierten“ Standards der evidenzbasierten Medizin und denen der Gesundheitsökonomie offensichtlich abweichen und hierfür auch plausible Erklärungen angeboten werden.

Nach dem IQWiG Vorgehen zur Zusatznutzenklassifizierung werden für jede Kategorie der Effektstärke (erheblich, beträchtlich, gering) **Schwellenwerte** definiert, und im Sinne einer verschobenen Hypothesengrenze muss ein 95%-Konfidenzintervall (für die Differenz zwischen der zu bewertenden Therapie und der zweckmäßigen Vergleichstherapie) diesen Schwellenwert unterschreiten, damit das Ergebnis entsprechend eingestuft wird. Für die Festlegung der Schwellenwerte wird ein zweischrittiges Verfahren eingeführt, wonach 1. eine Methode vorgestellt wird, die immer und unabhängig von der betrachteten Indikation und der Zielvariablen angewandt wird und 2. der klinische Schweregrad Berücksichtigung findet. Das IQWiG konstatiert als Regelfall für die verschobene Hypothese das Vorliegen zweier (pivotaler) Studien unter der Annahme, dass die Power aus zwei Einzelstudien zu den üblichen Hypothesen der Power der gemeinsamen (gepoolten) Analyse zu den verschobenen Hypothesen entspricht. Jedoch ist bereits diese Regelfallannahme in der Realität aus mehreren Gründen nicht gegeben. Regulatorisch werden nicht (struktur-) identische oder reproduzierte Studien u. a. aus folgenden Gründen eingefordert: a) Designparameter (Dosis, Applikationsform, initialer Schweregrad) bei vermuteter Effektmodifikation werden oft geändert; b) „verwandte“ Krankheiten werden mit je nur einer konfirmatorischen Studie bedient; c) niedrige Prävalenz und somit Patientenzahl bei seltenen Erkrankungen führen zu einer pivotalen Studie; d) Poolen nicht immer sinnvoll bzw. möglich aufgrund vorhandener Heterogenität ist (KI breiter); e) zwei homogene, etwa gleich große und gleichartige Studien eher einen Ausnahmefall darstellen; f) in der Onkologie beispielsweise regelhaft eine Studie durchgeführt wird und bei nachgewiesener Überlegenheit in Interimsanalysen oft Studien sogar vorzeitig beendet werden. Ferner erfolgt die entsprechende Fallzahlplanung mit realistischen Annahmen, also auch unter Berücksichtigung der weiter oben dargestellten Gegebenheiten, so dass die IQWiG Annahmen keinen Bestand haben. Aber auch die bis dato gesammelten Erfahrungen im Rahmen des AMNOG und der damit eingeführten frühen Nutzenbewertung nach § 35a SGB V zeigen, dass bei den erfolgten Bewertungen sowohl beim IQWiG als auch beim G-BA durchaus diese oft auf einer einzigen Studie fußten (von insgesamt 26 durch das IQWiG

durchgeführten Bewertungen wurden bei 9 Bewertungen keine relevanten Studien gefunden, 8 basierten auf 1 Studie, 6 auf 2 Studien und 3 auf 3 Studien, also erfolgte die Bewertung von etwa der Hälfte der Arzneimittel – 8 von 17 – auf Basis einer Studie). Laut § 5 (3) der AM-NutzenV sind für die Bewertung des Arzneimittels mit neuen Wirkstoffen grundsätzlich die Zulassungsstudien zugrunde zu legen, also die relevanten Zulassungsstudien, die eingereicht werden unabhängig von ihrer Anzahl. Somit ist die Regelannahme des IQWiG weder theoretisch begründet, noch praktisch bestätigt und es handelt sich um eine fiktive Planungsannahme. Im IQWiG Vorgehen zeigen sich des Weiteren Inkongruenzen zur AM-NutzenV hinsichtlich der Heilung und spürbare Linderung der Erkrankung als eigenständige Entitäten, die hier fehlen. Andererseits hat das IQWiG zu Recht die Lebensqualität ergänzt. Durch die inhaltlichen Abweichungen von der AM-NutzenV ergeben sich aber gravierende Implikationen, beispielsweise bei der Bewertung von Antibiotika oder Virustatika, die einen vom Gesetzgeber und der Exekutive möglicherweise nicht intendierten Einfluss auf die Bewertung von Arzneimitteln nehmen können. Ferner ist das Vorgehen mit implizierten Wertentscheidungen verknüpft, die nicht klar herausgearbeitet werden und die nach Auffassung des vfa nicht von einem Auftragsinstitut getroffen werden sollten, da dies hierzu nicht legitimiert ist. So werden die für die Herleitung der oberen **Konfidenzintervallschwellen** – dem Hypothesenshift unter der Annahme zweier pivotaler Studien folgend – eingehenden relativen Effektschätzer einer simplen Aufteilung in Mehrfachen von Sechsteln unterzogen (1/6~0,17, 2/6 ~0,33, 3/6~0,5, 4/6~0,67 und 5/6~0,83 sowie 6/6~1). Dies erfolgt laut IQWiG aus rein pragmatischen Gründen, nach Auffassung des vfa allerdings ohne wissenschaftlichen Hintergrund und eher willkürlich. Eine Validierung hierfür scheint zu fehlen. Hierbei verweist das IQWiG auf einen einzigen relativen Effektschätzer, nämlich den um HR=0,5 für die Mortalität, welcher als arbiträre Setzung eines Autorenkollektivs um Djulbegovic bei der Replik von 50 Jahren Onkologie in den USA neben der Originalautorendefinition von „Breakthroughs“ verwendet wurde und auch eindeutig als arbiträre Setzung in ihrer Publikation deklariert wird (Djulbegovic et al. 2008). Es handelt sich um eine explizit onkologische Betrachtung, ohne Differenzierung zur Wertigkeit der Mortalität innerhalb onkologischer Krankheitsbilder, geschweige denn bezogen auf andere Indikationen, die ohne gesetzte Annahmen keine Aussagen zu anderen Indikationen erlaubt. Bei 614 Studien zeigen sich nach der arbiträren Definition von Djulbegovic et al. 12 „Breakthroughs“ (~2%), wobei keine detaillierte Berücksichtigung

onkologischer Indikationen, Fallzahl, geplanter Power, absolute (Basis)Risiken, unterschiedliche Effektschätzer (HR vs OR vs RR) innerhalb dieser Studie erfolgt. Alle restlichen Effektschätzer zur Herleitung von Konfidenzintervallschwellen wurden ohne Verweis auf entsprechende Evidenz eingesetzt, und lassen vermuten, dass das IQWiG hier eine normative Entscheidung getroffen hat. Sie führen zu einer **impliziten Gewichtung** der Zielgrößen untereinander anhand einer willkürlichen mathematischen und nicht inhaltlichen Kategorisierung. So besteht zwischen den eingehenden Effektschätzern beispielsweise in der Zusatznutzenkategorie „erheblich“ für die Mortalität im Vergleich zur Lebensqualität ein Verhältnis von 3 zu 1 (RR/HR/OR 0,5 versus 0,17). Es drängt sich folgende Frage auf: Was könnte als Beispiel auch aus der Vergangenheit für einen klinischen Durchbruch nach IQWiG herhalten, das diesen Anforderungen genügt? Ferner bleibt völlig unklar, warum ein in der Onkologie als Durchbruch postulierter relativer Effektschätzer von 0,5 einem „erheblichen“ Ausmaß entsprechen sollte, und zudem warum dies neben der Onkologie auch in anderen Indikationsbereichen gelten sollte. Plakativ beschrieben wäre eine Reduktion der Mortalität von 100% auf 75% kein Durchbruch, wohingegen eine Reduktion der Mortalität von 1% auf 0.5% ein Durchbruch ist. Klinische Forschung zeichnet sich nicht ausschließlich durch „Durchbrüche“ aus, sondern profitiert häufig auch von kleineren, andauernden Fortschritten.

Auch die **Übertragbarkeit eines Effektschätzers** (sowie der Konfidenzintervalle) für RR auf andere Effektmaße wie HR oder OR kann sachlich nicht nachvollzogen werden. Ferner beschreibt der Entwurf, dass relative Maße gegenüber absoluter Risikoreduktion den Vorzug besitzen, dass es nicht der Benennung eines Auswertungszeitpunktes bedarf. Dies trifft nur auf das Hazard Ratio zu, nicht jedoch auf das Relative Risiko oder die Odds Ratio. Darüber hinaus stellt sich auch die Frage bei Betrachtungen der Zeit bis zum Ereigniseintritt, worin die Rechtfertigung besteht, für HR die gleichen Grenzen einzusetzen wie für RR.

Ferner werden für die Ausprägung erheblicher Zusatznutzen in den Endpunkten Schwerwiegende Symptome und Lebensqualität **Basisrisiken** $\geq$ 5% zumindest für eine der Vergleichsgruppen innerhalb der Studie vom IQWiG definiert, um in diese Zusatznutzenklasse überhaupt zu gelangen. Auch dies stellt nach Auffassung des vfa eine willkürliche Setzung dar, die nicht begründet erscheint. Plakativ beschrieben, würde dies dazu führen, dass beispielsweise eine Reduktion der Inzidenz schwerwiegender Symptome von 4,5% auf 0% als nicht erheblich eingestuft wird (da das Basisrisiko <5% ist), wohingegen

eine Reduktion von 5% auf 2.5% erheblich ist. Abschließend erachtet der vfa das vom IQWiG angewandte Vorgehen als eines, das keine allgemein in medizinischer Fachwelt akzeptierten Schwellenwerte zur Zusatznutzenklassifizierung generiert, die auch so keinen Bestand in den Standards der evidenzbasierten Medizin haben. Ferner erfolgt keine Berücksichtigung unterschiedlicher Indikationsgebiete mit deren jeweiligen (Basis)Risiken und der **Umgang mit nicht-dichotomen bzw. mit nicht in dichotome umwandelbare Zielvariablen** bleibt ungeklärt (bei ordinal kategoriell oder stetigen Variablen ist keine direkte Anwendung möglich). Darüber hinaus sind die Auswirkungen auf die Verschiebung der Hypothesengrenze unklar (Satz von Fieller). Nach Auffassung des vfa entziehen sich normative Fragen rein statistisch methodischen Antworten und Schwellenwerte können und sollten auch kontextabhängig sein. Nur indikationsspezifische Schwellenwerte können sicherstellen, dass unterschiedliche Ausgangswerte und evtl. auch absolute Basisrisiken Berücksichtigung finden können. Ein „one size fits all“ erscheint nicht möglich. Der vfa sieht im IQWiG Vorgehen a) ein Legitimierungsproblem, weil es normative Setzungen enthält und priorisiert; b) ein Eigenkonstrukt, dass nicht etabliert ist, weil es keinen internationalen Standard der evidenzbasierten Medizin darstellt, c) ein gravierendes Realisierungsproblem, weil die getroffenen Annahmen eben nicht den Regelfall widerspiegeln, noch die klinische oder regulatorische Realität abbilden.

Das Methodenpapier beschreibt die Vor- und Nachteile der relativen Effektschätzer. Als einen Nachteil der Verwendung von absoluten Risikoreduktionen als Effektmaß zur Operationalisierung hinsichtlich der Feststellung des Ausmaßes des Zusatznutzens bezeichnet das IQWiG die Notwendigkeit einer genauen Zeitpunktdefinition, bei dem die absolute Risikoreduktion bestimmt wird, sofern es dazu keine allgemein akzeptierten Festlegungen gibt (beispielsweise 30-Tage-Mortalität bei Myokardinfarkt). Allerdings benötigt man für die relativen Effektmaße RR und OR ebenso wie für Maße der absoluten Risikoreduktion einen genauen Zeitpunkt, bei dem diese Maße bestimmt werden sollen. Lediglich das HR hat diese Einschränkung nicht. Daher ist aus Sicht des vfa dieses Argument nicht schlüssig und kann auch nicht als Nachteil der Betrachtung absoluter Risikoreduktionen herangezogen werden. Unbeschadet der statistischen Eigenschaften wird nicht berücksichtigt, dass die veränderte Verfahrensordnung des G-BA die standardmäßige Darstellung der relativen Effektmaße OR und RR auch die Absolute Risikoreduktion ARR vorsieht. Zum anderen bestätigt das IQWiG implizit die oben dargestellte Forderung nach

einer indikationsspezifischen Betrachtung als Alternative zu starren Vorgaben zu Konfidenzintervallen: „Würde nun beispielsweise eine mindestens 20%ige absolute Risikoreduktion als wesentliche therapeutische Verbesserung definiert, so wäre (für diese beispielhafte Forderung) bei Erkrankungen mit (langfristigen) Überlebensraten >80% grundsätzlich kein erheblicher Zusatznutzen (für den entsprechenden Endpunkt) mehr darstellbar“ (IQWiG Methodenentwurf S. 31f.). Der grundlegende Fehler in dieser Betrachtung liegt nach Auffassung des vfa nicht an der Bewertung der absoluten Risikoreduktion an sich, sondern der indikationsübergreifenden unsachgemäßen Festlegung eines einheitlichen Wertes der für den Effektschätzer der absoluten Risikoreduktion als Definition einer „wesentlichen therapeutischen Verbesserung“.

Zu dem vorgeschlagenen Vorgehen des IQWiG zur Dichotomisierung bei stetigen oder kategorialen Variablen zwecks Überführung in relative Effektschätzer sei noch hier angemerkt, dass die Dichotomisierung vom eingesetzten Schwellenwert sehr stark abhängt. Des Weiteren ist dieses Vorgehen so nicht in der wissenschaftlichen Literatur beschrieben und beinhaltet u. a. undefinierte Approximationsverfahren. Ferner werden „etablierte“ Responderkriterien nicht näher erläutert. Mit der Dichotomisierung geht auch ein entsprechender Informationsverlust der betrachteten Variablen einher. Das vorgeschlagene Vorgehen führt dazu, dass die Zusatznutzenkategorie „nicht quantifizierbar“ aufgrund der Unmöglichkeit einer Überführung in ein relatives Effektmaß bemüht werden muss. Dies ist dem IQWiG Vorgehen inhärent und wird nicht durch das Fehlen einer Quantifizierung geschweige denn eines Zusatznutzens bedingt. Das IQWiG hat in der Vergangenheit bei stetigen Variablen u. a. das Konzept der standardisierten Mittelwertdifferenz angewendet. Somit scheitert nach Auffassung des vfa das IQWiG mit seinem Vorgehen auch für die vom IQWiG unter sonstige Zielgrößen bezeichneten Operationalisierungen von Endpunkten, weil es hierfür keine zielführende Lösung innerhalb seines Ansatzes zu finden scheint. Der vfa ist der Auffassung, dass ein Klassifizierungsschema für den Zusatznutzen nach einem spezifischen Algorithmus sofern es normative Urteile enthält, von einem hierfür legitimierten Entscheidungsträger verabschiedet werden müsste. Des Weiteren sollte es nicht rigide Setzungen bedienen, sondern entsprechende Anpassungen in Abhängigkeit von der Indikation und dem Schweregrad ermöglichen und bei Gewichtungen der Zielgrößen nicht implizit, sondern explizit und transparent unter Einbeziehung wenn möglich von Patientenpräferenzen bzw. durch institutionalisierte externe Gremien, die allerdings ihre

Entscheidung transparent machen, beispielsweise mittels der in Abschnitt 3.1.5 genannten Verfahren AHP und CA als Orientierung operieren. Dies vor Allem auch deswegen, weil das im Methodenentwurf des IQWiG enthaltene Vorgehen mit seinen Annahmen die Endpunkte letztlich in allen Indikationsgebieten gleich gewichtet. Schwerwiegende Komplikationen dürften allerdings bei potenziell letalen Erkrankungen wie in der Onkologie durch Patienten eher akzeptiert werden als bei anderen, weniger schwerwiegenden Erkrankungen. Ein identischer Zugewinn an Überlebensdauer wird von Patienten unterschiedlich bewertet, je nachdem wie die lang die Überlebenszeit überhaupt ist. Es kann somit kaum davon ausgegangen werden, dass die Bewertung von Nutzen und Schaden nicht durch das Krankheitsbild beeinflusst wird. Dies spricht gegen eine schematische, indikationsübergreifende Bewertung.

E) 7.3.8 Meta-Analysen

Dieser Abschnitt wurde in der Vergangenheit von hierfür berufenen Institutionen in seiner Fassung in der Methodenversion 4.0 kommentiert. So liegen beispielsweise Kommentare hierzu von Oliver Kuß aus Halle (Kuß 2012) im Auftrag der GMDS und der IBS vor, die nicht in dem neuen Abschnitt 7.3.8 Eingang gefunden haben. Dort wird Bezug genommen auf die die Nachteile einer routinemäßigen Anwendung der Standardmodelle mit zufälligen und festen Effekten und Methoden zu deren Verbesserung vorgeschlagen. In diesem Zusammenhang werden u. a. logistische Regressionsmodelle für korrelierte Beobachtungen genannt: gemischte Modelle (PQL, numerische lntegration mit Gauß-Quadratur), marginale Modelle (GEE), konditionale Modelle (Cox'sche Partial Likelihood), Modelle mit geschlossener Likelihood-Funktion (Beta-binomiale Regressionsmodelle) oder auch Modelle mit festem Studieneffekt (Standardmodell der logistischen Regression mit dem Studieneffekt als kategorieller Kovariable). Als Vorteil dieser Modelle gilt, dass sie die Annahme der festen und bekannten (die aber dann doch aus den Studien geschätzt werden müssen) Gewichte in den Standardverfahren vermeiden. Diese Modelle hätten den Vorteil, dass nicht notwendigerweise nur ein zufälliger Therapieeffekt angenommen werden muss, sondern darüber hinaus auch studienspezifische Basisrisiken oder die Interaktion von Studien und Behandlung als zufällig angenommen werden können. Der vfa schließt sich diesen Ausführungen an.

Bezüglich der Subgruppenanalysen im Rahmen von Meta-Analysen bleibt festzuhalten, dass die Verwendung von p-Werten (alpha = 0,2) zur Entscheidung hinsichtlich der Zusammenfassung von Subgruppen nicht immer adäquat erscheint, da die Fallzahl einen Einfluss auf den p-Wert hat. Darüber hinaus bleibt unklar, wie die Zusammenfassung bei zahlreichen Subgruppen und daher zahlreichen paarweisen Vergleichen wiederum aussehen soll.

Literaturliste:

Richter EA. Disease Management: Diabetes mellitus Typ 2 (II): „Wichtig ist die Blutdrucksenkung“. Deutsches Ärzteblatt 2002, 99(14): A 1632.

Schmieder RE. Hypertoniebedingte Endorganschäden. Deutsches Ärzteblatt 2010, 107(49): 866 -874.

Atkins D, Best D, Briss PA, Eccles MP, Falck-Ytter Y, Flottorp S et al. Grading quality of evidence and strength of recommendations. BMJ 2004, 328(7454): 1490.

EMA (European Medicines Agency). Points to consider on application with: 1. meta-analyses; 2. one pivotal study [online]. 31.05.2001 [Zugriff: 22.05.2013]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf

Dintsios CM. Patientengewichtete Endpunkte: Conjoint Analyse (CA) vs. Analytic Hierarchy Process (AHP). Monitor Versorgungsforschung 2013, Kongress-Special 01: 16-20.

Liberatore MJ, Nydick RL. The analytic hierarchy process in medical and health care decision making: A literature review. European Journal of Operational Research 2008, 189: 194–207.

Danner M, Hummel JM, Volz F, van Manen JG, Wiegard B, Dintsios CM, Bastian H, Gerber A, IJzerman MJ. Integrating patients’ views into health technology assessment: Analytic hierarchy process (AHP) as a method to elicit patient preferences. Int J Technol Assess Health Care 2011, 27(4): 369-375.

Hummel M, Volz F, van Manen J, Danner M, Dintsios CM, IJzerman M, Gerber A. Using the Analytic Hierarchy Process to elicit patient preferences: Prioritizing multiple outcome measures of antidepressant drug treatment. The Patient 2012, 5(4): 1-13.

Saaty TL. Decision making with the analytic hierarchy process. Int. J. Services Sciences 2008, 1(1): 83-98.

Ishizaka A., Labib A. Review of the main developments in the analytic hierarchy process, Expert Systems with Applications 2011, 38(11), 14336-14345.

Kuß O. Kommentare zum Abschnitt 7.3.8 Meta-Analysen (S. 137- 141). Gemeinsame Stellungnahme von GMDS und IBS-DR [online] 02.02.2012 (Zugriff 22.05.2013]. URL: http://www.gmds.de/pdf/publikationen/stellungnahmen/120202_Kommentare_Meta_Analysen.pdf

A.1.24 – Vinzenzkrankenhaus Hannover gGmbH

Von: Schüttert, Dr. Jan Bernd [redacted] / 13.05.2013 10:10 Uhr

Sehr geehrte Damen und Herren,

ich möchte mich gerne an einer Diskussion über das Methodenpapier 4.0 beteiligen und schicke Ihnen deshalb mit dieser e-mail meine Fragen und Anmerkungen.

Zu meiner Person: ich bin klinisch tätiger Arzt (Facharzt für Innere Medizin und Kardiologie) in einem Krankenhaus. In der Vergangenheit habe ich pharmagesponserte Vorträge gehalten. Trotzdem halte ich mich selbst für kritisch, unabhängig und (soweit man das überhaupt sein kann) unvoreingenommen.

Mich stört bzw. vermisse ich eine Erklärung der Begriffe zum Zusatznutzen der Medikamente (i.e. erheblich, beträchtlich, gering).
Wenn ich es richtig verstehe, wird der Zusatznutzen aktuell von Ihnen anhand der Reduktion des rel. Risikos bestimmt. Dies wird meiner Ansicht nach dem Problem bzw. der Fragestellung Ausmaß des Zusatznutzens nicht gerecht. Zum Beispiel ist meiner Ansicht nach eine Senkung der Mortalität über eine Zeitraum von 10 Jahren von 1 auf 0,75% (RR 0,25%)klinisch betrachtet völlig irrelevant.

Ferner verweisen Sie auf das dazugehörige Gesetz. Auch hier fehlen aber konkrete Angaben. Z. B. steht im Gesetz: „Ein erheblicher Zusatznutzen liegt vor, wenn eine…große Verbesserung des therapierelevanten Nutzens…erreicht wird, insbesondere… eine erhebliche Verlängerung der Lebensdauer…".
Wie viele Tage/Wochen/Monate/Jahre Lebensverlängerung ist denn „erheblich"?
Wer entscheidet das an Hand welcher Kriterien?
Entscheiden das Juristen (in Kommentaren), Richter (Prozesse), Krankenkassen, Pharmindustrie oder Ärzte?

Das Gleiche gilt für die übrigen Attribute zum Ausmaß des Zusatznutzens.

Vielleicht können Sie mich aufklären, vielleicht sind meine Kommentare/Fragen hilfreich.

Über eine Antwort von Ihnen würde ich mich sehr freuen.

Mit freundlichen Grüßen

Dr. J.B. Schüttert

Vinzenzkrankenhaus Hannover
gemeinnützige Gesellschaft mit beschränkter Haftung
Postfach 71 02 70
D- 30542 Hannover

Geschäftsführer: Michael Hartlage
Sitz der Gesellschaft: Hannover

[redacted]

A.2 – Stellungnahmen von Privatpersonen

A.2.1 – Meyer, Gabriele

Von: Meyer, Gabriele [redacted]
Gesendet: Sonntag, 19. Mai 2013 22:41
[redacted]
Betreff: Aktualisierung der Methoden 4.0 - Stellungnahme G Meyer

[redacted]
sehr geehrte Damen und Herren,

vielen Dank für die Möglichkeit, den ersten Teil der Aktualisierung des Methodenpapiers zu kommentieren. Es sind einige sehr wichtige Entscheidungen plausibel dargelegt und die methodischen Entscheidungen gut nachvollziehbar begründet.
Ich habe nur einen Vorschlag zur weiteren Optimierung. Unter „2.2.3 Review der Produkte des Instituts“ stellen Sie fest, dass die externen Reviews zu den Berichten nicht veröffentlicht werden. Das ist für mich nicht nachvollziehbar, denn zusätzliche Transparenz würde selbstverständlich geschaffen durch eine vollkommene Offenlegung der Gutachten und der konsekutiven Änderungen im Bericht. Die pre-publication history bei BioMed Central könnte hier Vorbild sein.
Des Weiteren hatte ich mir erhofft, dass Sie die Problemstellung Umgang mit komplexen Interventionen in das Methodenpapier aufnehmen. Vielleicht werden Sie diesen Methodendiskurs und Ihre Entscheidungen, wie Sie im (nicht-pharmakologischen Bereich) mit der systematischen Übersicht zu komplexen Interventionen umgehen wollen, in einem folgenden Teil der Aktualisierung berücksichtigen.

Mit freundlichen Grüßen
Gabriele Meyer

A.2.2 – Röhmel, Joachim

Stellungnahme von Joachim Röhmel zu Teilschritt 1: Entwurf der Änderungen der Allgemeinen Methoden 4.0 Aktualisierung einiger Abschnitte.

Bremen 18. Mai 2013

Kommentare und Kritiken sind entsprechend fortlaufenden Seiten im Entwurf sortiert.

A) Seite 2 letzter Abschnitt und Seite 4 4. Abschnitt: welche Öffentlichkeit ist gemeint?

B) Seite 8: Die Voraussetzung für Aussagen über das Fehlen eines (Zusatz-)Nutzens bzw. Schadens sind gut begründete Definitionen von Irrelevanzbereichen (siehe Abschnitt 7.3.6).
Dieser Satz ist auch schon in Methoden 4.0 enthalten. Die Definition von Irrelevanzbereichen für (Zusatz)Nutzen kann man gelegentlich antreffen. Für Schaden jedoch sind solche Irrelevanzbereiche nur extrem selten, auch wegen der vielen Möglichkeiten, wie sich ein Schaden äußern kann. Wenn sich das IQWiG an seine eigenen Aussagen hielte, könnte es nur in extrem seltenen Fällen das Fehlen eines Schadens annehmen. In den allermeisten Fällen müsste mit der Ungewissheit über Fehlen/Nichtfehlen umgegangen werden.

C) Seite 9: geringe qualitative Ergebnissicherheit: wo bleiben die einarmigen, nicht vergleichenden (bzw. nur prä-post Vergleiche) Studien. Fallen die ganz aus der Bewertung heraus?

D) Seite 9: Im Fall homogener Ergebnisse, die sich sinnvoll poolen lassen, muss der gemeinsame Effektschätzer statistisch signifikant sein.

Dieser Satz hier ist an dieser Stelle nicht verständlich. Wozu muss der gemeinsame Schätzer statistisch signifikant sein? Wenn er nun nicht signifikant ist? Erst später wird die Beleglage detailliert erörtert.

E) Seite 9: Falls das Prädiktionsintervall zur Darstellung der Heterogenität in einer Meta-Analyse mit zufälligen Effekten (siehe Abschnitt 7.3.8) dargestellt wird und den Nulleffekt nicht überdeckt, liegen gleichgerichtete Effekte vor.

Dies ist eine Definition, die dem Augenschein widersprechen kann, d.h. es ist der Fall denkbar, dass nicht alle Studienergebnisse in die gleiche Richtung weisen, obwohl das Prädiktionsintervall den Nulleffekt nicht überdeckt. Ist das gewollt? Könnte man denn vielleicht plausibel machen, dass folgenden Bedingungen (etwa bei mindestens 4 Studien) gleichwertig sind?

a) Das Prädiktionsintervall überdeckt den Nulleffekt nicht
b) Die Effektschätzer von zwei oder mehr Studien zeigen in eine Richtung. Für diese „gerichteten“ Studien gelten alle folgenden Bedingungen:
   - Das Gesamtgewicht dieser Studien ist ≥ 80%.

- Mindestens 2 dieser Studien zeigen statistisch signifikante Ergebnisse.
- Mindestens 50 % des Gewichts dieser Studien basiert auf statistisch signifikanten Ergebnissen.

Kann man sagen, dass (bei 2 oder 3 Studien) „deutlich gleichgerichtet“ definiert wird durch „alle Studien müssen signifikant sein“ (in der gleichen Richtung). Wenn bei 3 Studien „Gleichgerichtet-Sein“ vorliegt, aber nicht alle Studien signifikant sind, so ist automatisch „mäßig gleichgerichtet“ das Prädikat?

Mich irritiert deutlich der Doppelpunkt in „Nicht alle drei Studien weisen statistisch signifikante Ergebnisse auf: Die gleichgerichteten Effekte sind mäßig gleichgerichtet.“ Üblicherweise kommt in der deutschen Sprache nach dem Doppelpunkt die Erklärung für das, was vor dem Doppelpunkt steht.

F) Seite 17, 2 Absatz: Vom Effektmaß relatives Risiko ausgehend werden Zähler und Nenner immer so gewählt, dass sich der Effekt (sofern vorhanden) als Wert < 1 realisiert. D. h. ein Effekt ist umso stärker, je niedriger der Wert ist.

Diese Festsetzung berücksichtigt die dem Quotienten innewohnende Asymmetrie nicht ausreichend. Eine Zunahme der Heilungsrate z.B. unter Therapie A auf 40 Prozent (gegenüber 30% unter B) ist formal auf zwei Weisen nach obigem Satz transformierbar:

a) Betrachte B/A

b) Betrachte das gegenereignis der „Nonresponder-Raten“: 60 % unter A und 70% unter B und berechne dafür A/B

Im Falle a) könnte man mit einem Verhältnis 30/40=0.75 rechnen, im Falle b) mit 60/70= 0.857

Man beachte jedoch, dass Fall a) aber nicht mehr als Risiko Reduktion zu interpretieren ist, zudem auch nicht als Änderung gegenüber der Kontroll-Gruppe, da diese ja nicht zum Nenner des Quotienten beiträgt. Außer formalen Gründen (z.B. Arbeitsaufwand der Bewertung wächst, da Beurteilungskriterien angepasst werden müssten) gibt es meiner Ansicht nach keinen überzeugenden Grund, nicht auch Effektzuwächse zu zulassen.

Im Übrigen wird die von vielen Seiten bereits in der Vergangenheit geäußerte Kritik an der ausschließlichen Verwendung des relativen Risikos im neuen Entwurf so berücksichtigt, als hätte es diese Kritik nicht gegeben.

G) Seite 17: Es werden folgende drei Kategorien für die Qualität der Zielgröße gebildet:

☐ Gesamtmortalität

☐ Schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen sowie gesundheitsbezogene Lebensqualität

☐ Nicht schwerwiegende (bzw. nicht schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen

Zum einen handelt es sich weniger um „Qualität“ sondern mehr um „Schweregrad“ und zum zweiten lässt sich die Vielzahl unterschiedlicher Erkrankungen schwerlich in ein „Prokrustesbett“ von 3 Kategorien pressen. Hier sehe ich deutlich weiteren Diskussionsbedarf, in welchen aktiv die medizinischen Fachgesellschaften einbezogen werden müssten.

H) Seite 18/19: Bei den festgelegten Schwellen für das Erreichen von Effektstärken vermisse ich die sonst in dem Entwurf so gebräuchliche Formel „in der Regel“. Jetzt heißt es eher „das Konfidenzintervall muss vollständig unterhalb…“oder „die Schwellenwerte sind zu unterschreiten…“. So wie zurzeit formuliert könnte man Entscheidungen über Nutzen von Arzneimitteln nach kurzer Einübung – salopp formuliert - auch im „Kindergarten“ erledigen lassen. Das würde den Youngstern sicher Spaß machen und zudem das mathematische Verständnis zu Zahlen schulen.

Jedoch auch das Hinzufügen von „in der Regel“ würde das vorgeschlagene System nicht wirklich der Realität näher bringen. Eher schon das Einführen von Kern- und Graubereichen, ähnlich wie sie bereits vom IQWiG in der Vergangenheit bei der Diskussion zur klinischen Relevanz ins Gespräch gebracht worden ist: warum sollte bei Mortalität eine Reduktion des Risikos mit einer oberen Konfidenzgrenze von 0.86 nicht mehr „erheblich“ sein? Ist doch kontextabhängig! Richtig ist, dass in der Nutzenbewertung gemäß §35a SGB V eine Abstufung des Ausmaßes der Effekte vorgesehen ist. Tatsächlich sind Konfidenzintervalle und zugehörige Schwellenwerte auch aus meiner Sicht ***eine*** naheliegende biometrische Methode, um die Vorgaben des Gesetzes um zusetzen. Jedoch verlangt die der Gegenstand (Behandlung, Linderung und Vermeidung von Krankheiten sowie Entwicklung effektiver Arzneimittel) ein hohes Maß an Flexibilität, Diskussion und Mitbestimmung. Die Zeit ist Vergangenheit, in der mit Festsetzungen wie etwa das 5% Signifikanz-Niveau bei statistischen Tests ein hoher Akzeptanzgrad erreichbar ist. Im Übrigen ist es interessant zu sehen, dass der G-BA seine Entscheidungen zum Ausmaß „erheblich“, „beträchtlich“ oder „gering“ bisher nicht auf diesem Schema basiert, obwohl das IQWiG dieses Schema bereits seit der ersten Veröffentlichung im Bericht zu Ticagrelor (A11-02) anwendet.

I) Seite 18: Verwendung von Ereignissen oder Gegenereignissen. Hier geht der Entwurf auf die bereits erwähnte Asymmetrie ein, überlässt aber die Wahl der Transformation nun zum ersten Mal einer „inhaltlichen“ Diskussion. Gut so, nur bitte nicht nur an einer eher nebensächlichen Angelegenheit.

Seite 19: Das Konzept der alleinigen Betrachtung relativer Risiken wird vollends fragwürdig bei den Ausführungen zu stetigen bzw. quasi-stetigen Zielgrößen (vermutlich sind unter quasi-stetigen Zielgrößen z.B. die Ergebnisse von (multi-item) Fragebögen gemeint). Für viele stetige und quasi-stetige Zielvariable gibt es gar keine bisher allgemein akzeptierte oder

validierte oder etablierte Responder Definitionen. Hier würde sich eine weitere Diskrepanz zu Entscheidungen der Zulassung auftun, wo aus bekannten Gründen (z.B. Verlust an Information) Responder-Analysen basierend auf stetigen und quasi-stetigen Zielvariablen in der Regel nur zulassungsunterstützend angefordert werden. Bitteschön, national wo nötig, international wo möglich!

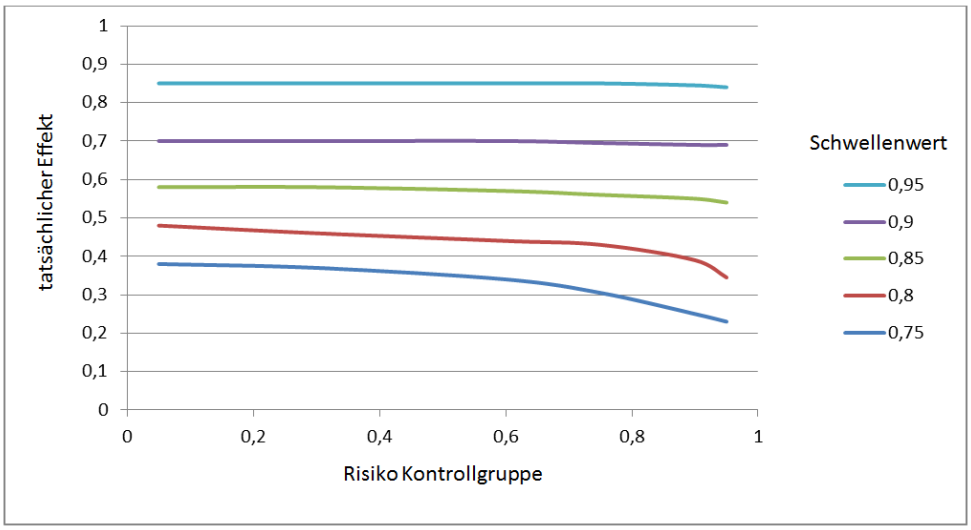
J) Seite 20: Bei Meta-Analysen ist die Verwendung von Prädiktionsintervallen neu. Diese Methode bekommt ein starkes Gewicht bei der Vergabe von den Prädikaten „Beleg", „Hinweis" und „Anhaltspunkt". Mögliche Diskrepanzen beim Sprung von 3 auf 4 Studien in einer Meta-Analyse, aber auch die abgestufte Verfahrensweise zuerst nur die qualitativ hoch stehenden Studien in eine Meta-Analyse einzubinden müssten erst noch ausgeräumt werden.

K) Seite 23 Untergruppen-Analysen. Während im Allgemeinen post hoc durchgeführte Subgruppenanalysen auf Studienebene kritisch zu interpretieren sind, ist man in einer systematischen Übersicht dennoch auf die Verwendung der Ergebnisse solcher Analysen auf Studienebene angewiesen, wenn im Rahmen der systematischen Übersicht genau diese Subgruppen untersucht werden sollen.

Tatsächlich wird durch Entscheidungen des G-BA (entweder wegen Problemen bei der Findung der zweckmäßigen Vergleichstherapie (ZVT) oder nach akribischem Lesen des Zulassungstextes) vom IQWiG verlangt, post-hoc definierte Untergruppen zu analysieren. Es ist wohl zu viel erwartet, dass dann solche post-hoc Untergruppen die verdiente kritische Würdigung erfahren würden. Hier wird einmal mehr verfahren nach dem Motto: „eine Regel gilt für die pharmazeutischen Sponsoren, eine andere Regel für die Zusatznutzen Bewertung". Eine solche Situation des Verlustes von Planbarkeit klinischer Studien hat es in Deutschland zuletzt in den 70ger und 80ger Jahren gegeben, wo pharmazeutische Hersteller erst mit den fertigen Dossiers zur Zulassung kamen und Ihnen dann klar gemacht wurde, dass die vorgelegten Dossiers nicht zur Zulassung reichen würden. Die damaligen Ereignisse fallen vielleicht nur zufällig zeitgleich mit einem von mir so empfundenen Niedergang der pharmazeutischen Industrie in Deutschland zusammen.

L) Seite 26 Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzen: Zu der Frage der Sinnhaftigkeit und Praktikabilität der Schwellenwerte habe ich bereits Stellung genommen. Ebenso habe ich öffentlich die im Anhang zum Bericht A11-02 genannten Ausführungen zum Ausmaß kritisiert. Leider ist wieder - selbst unter den Annahmen, die im Entwurf bei der Erstellung der Abbildung NA1 (Tatsächliche Effekte in Abhängigkeit des Basisrisikos) benutzt worden sind - eine von mir angeregte und im Entwurf genannte Diskussion fort zusetzen.

Abbildung NA1 Tatsächliche Effekte in Abhängigkeit des Basisrisikos



Ich war überrascht, dass eine in SAS - wie mir scheint – nicht validierte und aus meiner Sicht irreführende Formel zu verzerrten Ergebnissen geführt hat. Auch für die nun korrigierten Ergebnisse habe ich meine Zweifel zu ihrer Richtigkeit. Im Gegensatz zum Entwurf war eine Simulation dazu nicht notwendig. Unterschiede zwischen Entwurf und meinen Berechnungen kommen vermutlich durch die in der Simulation verwendeten Algorithmen zur Berechnung von Konfidenzintervallgrenzen. Erhebliche Abweichungen ergeben sich insbesondere für die niedrigen Grenzen 0.85, 0.8 und 0.75. Gelegentlich wurde geäußert, dass meine Berechnungen auf exakten Methoden beruhen und daher mit Ergebnissen aus dem Anhang zu Ticagrelor (A11-02) nicht vergleichbar seien. Tatsächlich könnte ich dies ohne großen Aufwand tun. Wegen der Vergleichbarkeit mit den Methoden im Entwurf habe ich bisher darauf verzichtet. Meine Berechnungen basieren auf einer Arbeit von Farrington und Manning (test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unit relative risk. Statistics in Medicine 9, 1990, 1447-1454) zu approximativen Tests bei verschobenen Hypothesen. Diese Formeln basieren wiederum auf dem „restricted maximum likelihod principle“ (REML). Farrington und Manning (FM) haben diese Methode nicht erfunden, jedoch sind numerische Aspekte in der genannten Arbeit sehr detailliert dargestellt, und sie werden von vielen Kollegen in biometrischen Umfeld bei Fallzahlberechnungen eingesetzt. Wegen der bekannten Dualität zwischen Konfidenzintervallen und statistischen Tests mit verschobenen Hypothesen ist die Arbeit von Relevanz.

Nennen wir den Algorithmus N=FM(alpha, beta, p, R0, S0), welcher bei gegebenem Signifikanzniveau alpha (üblicherweise 2.5% einseitig), Power (1-beta) (üblicherweise 90%), Basisrisiko p (in der Kontrollgruppe, zwischen 5% und 95% liegend), gewünschter Risiko Reduktion *R0* (0.167, 0.333, 0.500, 0.667, 0.833) und Ausmaßschwelle S0 die Fallzahl N ermittelt (gleich groß in beiden Gruppen, der Algorithmus kann aber auch ungleich große Gruppen behandeln). Die zu *R0* korrespondierenden Schwellen sind *S* (0.75, 0.80, 0.85, 0.90 und 0.95)

Dem Entwurf folgend beginnen wir mit einer Fallzahlschätzung für eine konventionelle Studie, die für eine Nullhypothese mit dem (relativen) Nulleffekt S0=1 geplant ist

N= FM(alpha, beta, p, *R0*, S0=1)

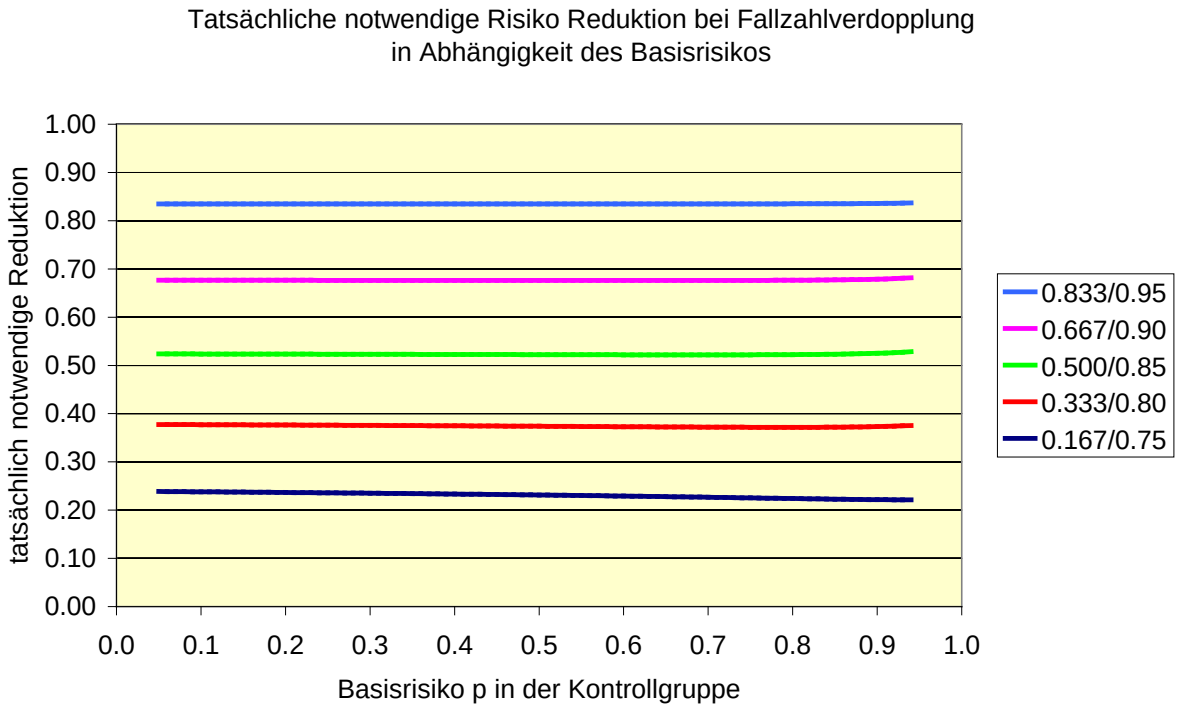
Dann ist es durch Invertierung nicht weiter schwierig, die Risiko Reduktion R1 (*R0* < R1 < *S*) zu ermitteln, die (bei gegebenem Risiko p in der Kontrollgruppe) bei Verdopplung der Fallzahl die folgende Beziehung erfüllt:

2N=FM(alpha, beta, p, R1, *S*),

wobei (*R0, S*) der Reihe nach die Paare

(0.833, 0.95), (0.667, 0.90), (0.500, 0.85), (0.333, 0.80) und (0.167, 0.75) durchläuft.

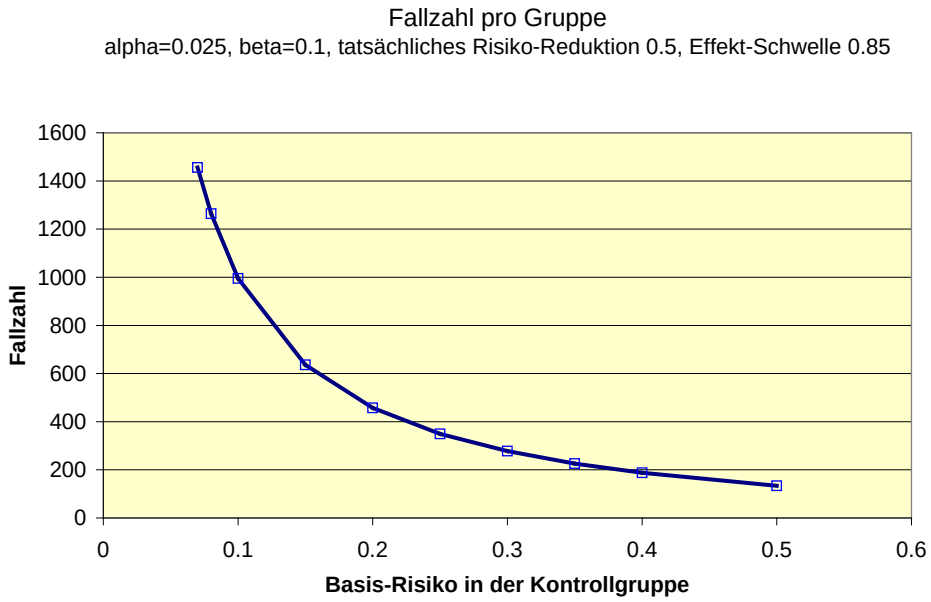
Daraus ist die unten stehende Graphik ermittelt (mit Basis Risiko p auf der x-Achse variierend zwischen 0.05 und 0.95), die viel eher als die im Entwurf publizierte Graphik ein vom Basis-Risiko unabhängiges Verhältnis von R0 zu R1 anzeigt. Von daher wäre eine Abstufung der erwünschten Risiko Reduktionen in Stufen von 1/6 bis 5/6 eigentlich auch überflüssig, weil sich der Übergang kontinuierlich und (fast) unabhängig vom Basis Risiko gestaltet, also die Relation der tatsächlichen Risiko Reduktion zur erwünschten Risiko-Reduktion bei Fallzahl Verdopplung überschaubar bleibt.



Die Diskrepanzen sind teilweise erheblich. Es besteht Diskussionsbedarf.

M) Planbarkeit und Fallzahlplanung insbesondere. Nicht unerheblich werden die „Planungsaspekte“ über gewünschte und geforderte Effekte von Ergebnissen der dazu

benötigten Fallzahlen beeinflusst. Dieses Thema ist bisher kaum untersucht worden. Aus meiner Sicht zeigt sich, dass die rigorose Festsetzung von Schwellenwerten für die Kategorisierung des Ausmaßes von Effekten unabhängig von der Indikation den Aufwand für die Durchführung von Arzneimittelstudien in einigen Indikationen erheblich erhöht. Zum Beispiel wird dazu im Entwurf des IQWiG Bezug genommen auf eine Arbeit von Djulbegovic (Djulbegovic B, Kumar A, Soares HP, Hozo I, Bepler G, Clarke M et al. Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006. Arch Intern Med 2008; 168(6): 632-642), der ein relatives Risiko von 0.5 als Anforderung für einen „Durchbruch" postuliert. Dieser Effekt wurde vom IQWiG dann als Effekt erheblichen Ausmaßes für die Zielgröße Gesamtmortalität im Anhang zum Bericht A11-02 verankert. Beim Lesen der Arbeit von Djulbegovic fällt auf, dass diese Untersuchung sich nur mit onkologischen Studien befasst (siehe auch Titel), und die Sterblichkeit im Beobachtungszeitraum erheblich ist. Das ist von Bedeutung, da – wie in der unten stehenden Abbildung gezeigt wird - die Fallzahl mit steigendem Basis-Risiko in der Kontrollgruppe erheblich absinkt und in den Bereich unter 200 Patienten pro Gruppe absinkt. Die mittlere Gesamtpatientenzahl in den Studien, die in Djulbegovic beschrieben werden, betrug 350.



In dieser Abbildung werden die benötigten Patientenzahlen angegeben (pro Gruppe), wenn ein Sponsor eine Studie mit dem Ziel durchführen wollte, einen Beleg für eine erheblichen Zusatznutzen bei Gesamtmortalität zu erreichen, und dabei von der Annahme ausgegangen wird, dass die zu prüfende Therapie eine 50%ige Reduktion der Mortalität erzielt.

Im kardiologischen Bereich (Basis-Risiko unter der Kontrollgruppe zwischen 5% und 10%) sind in der Vergangenheit Studien mit weitaus größeren Patientenzahlen durchgeführt worden (z.B. PLATO (NEJM 2009, 361(11), 1045-1057)) mit ca. 8000 Patienten pro Gruppe). Einzelstudien sind dann die Regel, eine Meta-Analyse die Ausnahme. Es sei daran erinnert, dass Studien bisher fast ausnahmslos im herkömmlichen Sinne geplant wurden, d.h. mit dem Ziel, einen Nulleffekt auszuschließen. Die Tatsache, dass dennoch große Patientenzahlen gebraucht wurden zeigt, dass die Studien für erheblich kleinere als 50%ige Risiko

Reduktionen geplant waren, sicherlich in der Erwartung, dass die geplanten Effekte als „von hoher klinischer Relevanz“ eingestuft werden würden. Tatsächlich ist mir keine Studie mit kardiovaskulärer Mortalität als Endpunkt bekannt, bei der die Planungsannahmen eine 50%ige Reduktion der Mortalität gegenüber einem aktiven Komparator vorsahen.

Zum Schluss noch ein Beispiel zu den Zahlenverhältnissen: Planung einer Studie mit Endpunkt Mortalität mit gewünschtem Effekt einer 50%igen Reduktion und Nachweis, dass das 95% Konfidenzintervall vollständig unter dem Schwellenwert von 0.85 liegt (Zusatznutzen erheblich), erfordert 1456 Patienten pro Gruppe, wenn das Basis-Risiko in der Kontrollgruppe 7% betragt; während eine Studie ebenfalls zum Nachweis eines erheblichen Zusatznutzens bei einem Endpunkt „Schwerwiegende Symptome“ (auch mit Basis-Risiko von 7%) und einer gewünschten Reduktion von 83% und Schwelle für das Konfidenzintervall von 0.75 „nur“ 345 Patienten pro Gruppe erfordert (alpha=2.5% einseitig, beta=10%). Der Begriff „gewünschte Effekte“ ist im Anhang zu Ticagrelor: Nutzenbewertung gemäß § 35a SGB V; Dossierbewertung; Auftrag A11-02 definiert und dient als Basis für die Ableitung der Schwellenwerte für das Ausmaß der Effekte. Könnte es also sein, dass die dort angegebenen gewünschten Effekte gar nicht als Grundlage für reale Studienplanung dienen sollen?

A.2.3 – Wink, Konrad

Prof. Dr. med. K. Wink
Medizinische Fakultät der Universität Freiburg
Arzt für Innere Medizin / Kardiologie
[redacted]

15. 5. 2013

Stellungnahme

zur

Überarbeitung Methodenpapier: IQWiG stellt Änderungen am Methodenpapier zur Diskussion

Auf den ersten Seiten wird das formale Vorgehen zur Berichterstellung dargestellt. Inhaltliche Fragen zur Bewertung finden sich in diesen Seiten nicht.

In 2.2.3 Review der Produkte des Instituts wird betont, dass insbesondere zum Ziel eine hohe wissenschaftliche Qualität der Produkte zu gewährleisten ist.

3.1.4 Bei der Endpunkt bezogenen Bewertung orientiert sich die Nutzenbewertung und Einschätzung der Stärke der Ergebnis(un)sicherheit am internationalen Standard der evidenzbasierten Medizin.
Dabei wurde nicht berücksichtigt, dass bei pivotalen Studien primäre, sekundäre, kombinierte Endpunkte und Surrogat-Endpunkte unterschieden werden. Entscheidende Bedeutung hat der primäre Endpunkt. Seine Veränderung wird festgelegt. Mit Hilfe des α- und ß-Fehlers wird daraus die Fallzahlbestimmung durchgeführt. Es wird damit eine Hypothese definiert, bei deren Erfüllung von einem aussagekräftigen Ergebnis ausgegangen wird, das in der Regel dann auch zur Zulassung des Medikaments führt.
Die Nichtbeachtung der Bedeutung des primären Endpunktes schwächt die Aussagekraft einer Studie. Man behandelt dann den primären Endpunkt wie einen sekundären Endpunkt, d.h. nicht hypothesenbeweisend sondern nur hypothesengenerierend.
Die Nutzung von Surrogatendpunkten ist sehr problematisch, da genügend Belege bestehen, die zeigen, dass sie nicht immer aussagekräftig für einen klinisch relevanten Endpunkt sind.
Die Berücksichtigung der primären Endpunkte als Hypothesenbeleg verbessert die Ergebnissicherheit. Eine hohe qualitative Ergebnissicherheit ergibt sich bei einer randomisierten Studie nur, wenn die Bedeutung des primären Endpunktes berücksichtigt wird. Eine pivotale randomisierte, kontrollierte Studie mit einem primären Endpunkt kann eine höhere Aussagekraft haben als die Zusammenfassung mehrerer Studien in einer Meta-Analyse, in der primäre Endpunkte nicht berücksichtigt werden.

Werden mehrere Studien zusammengefasst und in einer Meta-Analyse ausgewertet, muss berücksichtigt werden nach der Patientenzahl, ob die Methode nach der inversen Varianz, nach Peto oder nach Mantel-Haenszel erfolgt. Je nachdem, ob die Unterschiede nur innerhalb

der Studien oder zwischen den Studien auftreten, muss entweder das fixed- effect- oder das random-effect-Modell gewählt werden und wichtig ist auch nach der Zahl der Ereignisse, ob bei den Messparametern das relative Risiko, die Odds ratio oder die Risikodifferenz verwandt wird. Bei der Streuung hat man sich allgemein auf die 95%igen Konfidenzintervalle geeinigt. Da die verschiedenen Studien bei unterschiedlichen Patientenzahlen häufig divergente Ergebnisse, auch bei unterschiedlichen Messparametern, aufweisen, sind die verschiedenen Methoden, Modelle und Messparameter zu berücksichtigen.

3.1.5 Zusammenfassende Bewertung

Es ist sicher nicht unproblematisch, bei einer gleichzeitigen Würdigung von Nutzen und Schaden verschiedene patientenrelevante Endpunkte zu einem einzigen Maß zu aggregieren, da Nutzen und Schaden in Abhängigkeit vom Krankheitsbild aber auch von Seiten der Patienten sehr unterschiedlich empfunden werden können. Die Kriterien der Wirkung, der Nebenwirkungen, Lebensqualität sollten zunächst einmal nebeneinander dargestellt werden, bevor sie unterschiedlich bewertet werden.

3.3.3 Nutzenbewertung von Arzneimitteln gemäß §35a SGB V

1. Zugelassene Anwendungsgebiete
   Beim Vergleich verschiedener Studien sollte die Zulassung, die nach §25 AMG erfolgt, möglichst identisch sein.
   Dies ist eher nicht der Fall, da doch von den Herstellern Bestrebungen bestehen, Indikationsbereiche eher auszudehnen und damit andere Kollektive in die Studie einzubeziehen und Endpunkte zu wählen. Dies führt jedoch zu einer Heterogenität, die u.U. Vergleiche nicht mehr zulässt und damit auch Überlegenheiten, Äquivalenz oder Nicht-Unterlegenheit nicht beurteilt werden können.
   Auch muss darauf geachtet werden, dass nicht aus dem eingeschlossenen Kollektiv Substudien kreiert werden oder Subgruppen-Analysen stark bewertet werden.
   Das Aufteilen eines Gesamtkollektivs in einzelne Gruppen, die nicht mehr der ursprünglichen Planung und Festlegung einer konfirmatorischen Studie entsprechen, kann zu Verschiebungen der Ergebnisse (Simpson'sches Paradoxon), aber auch zu einer Minderung der Aussagekraft dieser Substudien führen, da es sich nicht mehr um gleiche und normale Verteilungen der Patienten handelt und nur noch eine Per Protokoll-Analyse bei Überlegenheitsstudien durchgeführt wird. Auch mit dem Regression-zum Mittelwert-Effekt muss gerechnet werden.
   Auch die Wahl der Vergleichstherapie kann aus den obigen Gründen schwierig werden.
   Die Anwendungen indirekter Vergleiche sind mit herabgesetzter Aussagekraft nur adjustiert möglich. Dabei ist jedoch die Heterogenität zu berücksichtigen, wobei allerdings nicht gesichert ist, ab welchem Ausmaß an Heterogenität man auf einen indirekten Vergleich verzichten sollte. Diese Heterogenität beruht auf verschiedenen Faktoren wie Zusammensetzung des Kollektivs, der Intervention, der Co-Intervention,

der Dauer der Studie, der Ausfälle und auch der Studienqualität. Auch beim adjustierten indirekten Vergleich muss etwa in 12% mit Diskrepanzen gerechnet werden.

4. Die Bestimmung eines Kollektivs, für das ein therapeutisch bedeutsamer Nutzen besteht, ist allenfalls als Hinweis bei Subgruppenanalysen zu entnehmen. Die Aussagekraft ist jedoch dabei deutlich reduziert und im Grunde können solche Subgruppenanalysen-Ergebnisse auch bei signifikantem Interaktionstest nur als Hypothesen verwendet werden.

   Bei den quantifizierbaren Effekten wird das Vorgehen mit binären Zielgrößen aufgeführt. In der medizinischen Biometrie werden allerdings Intervalldaten gegenüber Nominaldaten bevorzugt.
   Die Reduktion auf stetige binäre Endpunkte bedingt jedoch einen Güteverlust, Probleme bei der Interpretation, unterschiedliche biometrische Methoden zwischen den Zulasssungsstudien und bei der Auswertung der Studien zur Nutzenbewertung;
   eine höhere Fallzahl wird benötigt und die Power nimmt ab. Man müsste in Zukunft unterscheiden zwischen Studien zur Zulassung oder Nutzenbewertung.

   Bei den Kategorien für die Qualität der Zielgrößen wird eine Hierarchie, bestehend aus
   - Gesamtmortalität
   - Schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen, sowie gesundheitsbezogene Lebensqualität
   - Nicht-schwerwiegende (bzw. nicht-schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen

   unterschieden.

   Diese Graduierung hängt von den Krankheitsbildern und den Endpunkten ab. So kann z.B. die Gesamtmortalität von geringerer Bedeutung sein bei Patienten, die nicht an der untersuchten Krankheit sterben.
   Ein Beispiel hierfür könnte das Prostata-Ca sein, wobei die meisten Patienten wegen einer anderen Ursache außerhalb des Prostata-Ca sterben. Hier wäre z.B. das progressionsfreie Überleben ein wichtigerer Endpunkt. Auch schwerwiegende und nicht-schwerwiegende Symptome bzw. Nebenwirkungen hängen in Relation von der Schwere des Krankheitsbildes ab. Eine allgemeine Quantifizierung ist deshalb nicht möglich und muss für jedes Krankheitsbild getrennt festgelegt werden.
   Die Formel zur Festlegung des oberen Konfidenzintervalls ist strittig.

7.3.8 Meta-Analysen
Wie bereits vorne ausgeführt, muss man bei der Zusammenfassung von Studien zu einer Meta-Analyse davon ausgehen, dass die Patientenzahlen unterschiedlich sind, Studien mit gleich gerichteten und zufälligen Effekten und auch unterschiedlichen Messparametern bestehen. Es ist daher nützlich, alle Methoden-Modelle und Messparameter zu berücksichtigen, auch wenn die Unterschiede nicht immer groß sind.
Wenn sich aber deutliche Unterschiede ergeben, ist zu überlegen, ob die Durchführung einer Meta-Analyse sinnvoll ist.

Zur Problematik der Bewertung von Subgruppen-Analysen siehe vorne.
Unsicherheiten bleiben jedoch, wenn allein auf Grund von Subgruppen-Analysen ein Zusatznutzen festgelegt werden sollte.

Geringe Zahl von Ereignissen
Beim Ersatz von Nullzellen durch den Korrekturwert von 0,5 können sehr unterschiedliche Gesamtergebnisse entstehen. Es gibt Beispiele dafür, dass eine einzige Nullzelle, die durch 0,5 ersetzt wurde, zu einem signifikanten Ergebnis führte, das bei der Nullzellen-Berechnung nicht bestand.

Überlegungen zur Quantifizierung der klinischen Relevanz sind nicht ausgeführt. Das könnte aber auch bedeuten, dass die Einschätzungen der Betroffenen so unterschiedlich sind, dass sie nur grob biometrisch festgelegt werden können. Hier ist sicher in der Bewertung eine Zusammenarbeit mit den Betroffenen notwendig. Entsprechende Studien fehlen jedoch.

Keine Erwähnung finden auch pharmakologische Ergebnisse, die infolge des hier möglichen Reduktionismus stabilere Hinweise ergeben können als klinische Studien mit vielen Confounders.

Es erscheint nahezu unmöglich, retrospektiv aus Studien, die nicht hinsichtlich der Gewinnung eines Zusatznutzens geplant und durchgeführt wurden, sichere stabile Ergebnisse zu erzielen, die die Ermittlung eines Zusatznutzens ermöglichen. Es darf darauf hingewiesen werden, dass sehr ähnliche Studien, in kurzer Zeit nacheinander durchgeführt, zu unterschiedlichen Ergebnissen führen können. Es wird deshalb notwendig sein, für die Ermittlung eines Zusatznutzens prospektive Studien mit direktem Vergleich (head-to-head) durchzuführen.

Auch sollte man sich bei der Ermittlung eines Zusatznutzens aus retrospektiven Studien nicht zu weit vom internationalen Stand wissenschaftlicher Erkenntnisse entfernen und notfalls diese Auswertungen politisch begründen.

K. Wink