Gemeinsamer Bundesausschuss

Zusammenfassende Dokumentation

zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Verfahrensordnung: Änderung der Modulvorlage in der Anlage II zum 5. Kapitel

Vom 16. Dezember 2021

Inhalt

A. Tragende Gründe und Beschluss

1 Rechtsgrundlage

Der Gemeinsame Bundesausschuss (G-BA) hat gemäß § 91 Absatz 4 Satz 1 Nummer 1 SGB V eine Verfahrensordnung zu beschließen, in der er insbesondere methodische Anforderungen an die wissenschaftliche sektorenübergreifende Bewertung des Nutzens, der Notwendigkeit und der Wirtschaftlichkeit von Maßnahmen als Grundlage für Beschlüsse sowie die Anforderungen an den Nachweis der fachlichen Unabhängigkeit von Sachverständigen und anzuhörenden Stellen, die Art und Weise der Anhörung und deren Auswertung regelt. Die Verfahrensordnung bedarf gemäß § 91 Absatz 4 Satz 2 SGB V der Genehmigung des Bundesministeriums für Gesundheit. Mit Beschluss vom 20. Januar 2011 hat der G-BA ein 5. Kapitel in die Verfahrensordnung (VerfO) eingefügt, in dem das Nähere zum Verfahren über die Bewertung des Zusatznutzens von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V geregelt ist.

2 Eckpunkte der Entscheidung

Mit dem vorliegenden Beschluss werden Anpassungen der Anlage II.6 (Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen) zum 5. Kapitel der VerfO vorgenommen, welche durch Änderungen der methodischen Anforderungen an die Dossiererstellung in Verbindung mit dem bisherigen Vorgehen und den Erfahrungen des G-BA mit der Nutzenbewertung nach § 35a SGB V erforderlich geworden sind.

Die Änderungen betreffen in dem Abschnitt 4.3.1 (Ergebnisse randomisierter kontrollierter Studien mit dem zu bewertenden Arzneimittel) den Unterabschnitt 4.3.1.3.1 (<Endpunkt xxx> – RCT) der Anlage II.6 zum 5. Kapitel der VerfO. Diesbezüglich werden Konkretisierungen zur Ergebnisdarstellung von patientenberichteten Endpunkten vorgenommen, wie das Vorgehen zur Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen erfolgen soll.

Gemäß dem aktuellen methodischen Vorgehen des IQWiG (Methodenpapier 6.0, veröffentlicht am 5. November 2020) erachtet das IQWiG für patientenberichtete Endpunkte eine Responseschwelle für Responderanalysen von mindestens 15 % der Skalenspannweite eines Instrumentes (bei post hoc durchgeführten Analysen von genau 15 % der Skalenspannweite) als notwendig, um eine für Patienten spürbare Veränderung hinreichend sicher abzubilden.

Hintergrund ist, dass sich bei Responderanalysen auf Basis eines Responsekriteriums im Sinne einer individuellen Minimal important Difference (MID) vermehrt methodische Probleme offenbart haben.

Demnach zeigen systematische Zusammenstellungen empirisch ermittelter MIDs, dass zu einzelnen Instrumenten häufig eine Vielzahl von MIDs publiziert werden, die innerhalb eines

Erhebungsinstruments große Spannweiten haben können[1, 2, 3, 4]. Ursächlich hierfür können unter anderem die in den Studien eingesetzten unterschiedlichen Anker, Beobachtungsperioden oder analytischen Methoden sein[10, 5, 6]. Gleichzeitig ist eine anhand methodischer Qualitätskriterien begründete Auswahl empirisch ermittelter MIDs für die Nutzenbewertung derzeit nicht zu treffen[11, 7, 8].

Neben den methodischen Faktoren beruht ein anderer Teil der Variabilität von MIDs auf ihrer Abhängigkeit von Charakteristika der Patientenpopulation, in der das Instrument eingesetzt wird, sowie weiteren Kontextfaktoren. So können der Schweregrad der Erkrankung, die Art der eingesetzten Intervention oder die Frage, ob die Patientinnen und Patienten eine Verbesserung oder Verschlechterung ihrer Erkrankung erfahren, Einfluss auf die MID haben9. Der Umgang mit diesem Teil der Variabilität von MIDs ist ungeklärt.

Insgesamt gehen die genannten Limitationen bei Responderanalysen auf Basis eines Responsekriteriums im Sinne einer MID mit wesentlichen Unsicherheiten bei der Interpretation der klinischen Relevanz des beobachteten Effektes einher.

Im IQWiG-Methodenpapier (Methodenpapier 6.0, veröffentlicht am 5. November 2020) wurde ein Wert von 15 % der Spannweite der jeweiligen Skalen als plausibler Schwellenwert für eine hinreichend sicher spürbare Veränderung empirisch gestützt hergeleitet.

Die mit dem vorliegenden Beschluss vorgenommene Anpassung der Anlage II.6 zum 5. Kapitel der VerfO soll daher sicherstellen, dass für Responderanalysen im Rahmen der Nutzenbewertung geeignete Responseschwellen eingesetzt werden, die eine für die Patientinnen und Patienten hinreichend sicher spürbare Veränderungen abbilden. Zudem wird so die Gefahr einer ergebnisgesteuerten Berichterstattung minimiert, so dass diesbezügliche Unsicherheiten bei der Interpretation der klinischen Relevanz des beobachteten Effektes verhindert werden sollen. Mit diesen Änderungen der methodischen Anforderungen an die Dossiererstellung in Verbindung mit dem bisherigen Vorgehen und den Erfahrungen des G-BA mit der Nutzenbewertung nach § 35a SGB V sollen zudem

1 Carrasco-Labra A, Devji T, Qasim A, Phillips M, Devasenapathy N, Zeraatkar D et al. Interpretation of patient-reported outcome measures: an inventory of over 3000 minimally important difference estimates and an assessment of their credibility. Cochrane Database Syst Rev 2018; (9 Suppl 1): 135-136.

2 Çelik D, Çoban Ö, Kılıçoğlu Ö. Minimal clinically important difference of commonly used hip-, knee-, foot-, and ankle-specific questionnaires: a systematic review. J Clin Epidemiol 2019; 113: 44-57.

3 Hao Q, Devji T, Zeraatkar D, Wang Y, Qasim A, Siemieniuk RAC et al. Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation. BMJ Open 2019; 9(2): e028777.

4 Nordin A, Taft C, Lundgren-Nilsson A, Dencker A. Minimal important differences for fatigue patient reported outcome measures: a systematic review. BMC Med Res Methodol 2016; 16: 62.

5 Devji T, Guyatt GH, Lytvyn L, Brignardello-Petersen R, Foroutan F, Sadeghirad B et al. Application of minimal important differences in degenerative knee disease outcomes: a systematic review and case study to inform BMJ Rapid Recommendations. BMJ Open 2017; 7(5): e015587.

6 Ousmen A, Touraine C, Deliu N, Cottone F, Bonnetain F, Efficace F et al. Distribution- and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review. Health Qual Life Outcomes 2018; 16(1): 228

7 Devji T, Carrasco-Labra A, Lytvyn L, Johnston B, Ebrahim S, Furukawa T et al. A new tool to measure credibility of studies determining minimally important difference estimates. Cochrane Database Syst Rev 2017; (9 Suppl 1): 58.

8 Johnston BC, Ebrahim S, Carrasco-Labra A, Furukawa TA, Patrick DL, Crawford MW et al. Minimally important difference estimates and methods: a protocol. BMJ Open 2015; 5(10): e007953.

Unsicherheiten der pharmazeutischen Unternehmer in der Dossiererstellung im Hinblick auf die Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen vermieden werden. Gleichzeitig sollen mit den Änderungen in der Modulvorlage Lücken in der Ergebnisdarstellung zu patientenberichteten Endpunkten, welche teilweise erst im Rahmen des Stellungnahmeverfahrens aufgeklärt werden können, vermieden werden.

Mit dem vorliegenden Beschluss zur Anpassung der Anlage II.6 zum 5. Kapitel der VerfO wird der Unterabschnitt 4.3.1.3.1 (<Endpunkt xxx> – RCT) daher wie folgt geändert:

Es wird ergänzt, wie das Vorgehen zur Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen im Rahmen der Dossiererstellung erfolgen soll:

1. Falls in einer Studie Responderanalysen unter Verwendung einer MID präspezifiziert sind und das Responsekriterium mindestens 15 % der Skalenspannweite des verwendeten Erhebungsinstruments entspricht, sind diese Responderanalysen des Responsekriteriums für die Bewertung darzustellen.
2. Falls präspezifiziert Responsekriterien im Sinne einer MID unterhalb von 15 % der Skalenspannweite liegen, bestehen in diesen Fällen und solchen, in denen gar keine Responsekriterien präspezifiziert wurden, aber stattdessen Analysen kontinuierlicher Daten zur Verfügung stehen, verschiedene Möglichkeiten. Entweder können post hoc spezifizierte Analysen mit einem Responsekriterium von genau 15 % der Skalenspannweite dargestellt werden. Alternativ können Analysen der kontinuierlichen Daten dargestellt werden, für die Relevanzbewertung ist dabei auf ein allgemeines statistisches Maß in Form von standardisierten Mittelwertdifferenzen (SMDs, in Form von Hedges' g) zurückzugreifen. Dabei ist eine Irrelevanzschwelle als Intervall von - 0,2 bis 0,2 zu verwenden: Liegt das zum Effektschätzer korrespondierende Konfidenzintervall vollständig außerhalb dieses Irrelevanzbereichs, wird davon ausgegangen, dass die Effektstärke nicht in einem sicher irrelevanten Bereich liegt. Dies soll gewährleisten, dass der Effekt hinreichend sicher mindestens als klein angesehen werden kann.
3. Liegen sowohl geeignete Responderanalysen (Responsekriterium präspezifiziert mindestens 15 % der Skalenspannweite oder post hoc genau 15 % der Skalenspannweite) als auch Analysen stetiger Daten vor, sind die Responderanalysen darzustellen.

Der G-BA hat in seiner Sitzung am 17. Juni 2021 die Einleitung eines fakultativen Stellungnahmeverfahrens gemäß 1. Kapitel § 8 Absatz 2 Satz 1 Buchstabe b VerfO zur Änderung der VerfO beschlossen.

Im Anschluss an das schriftliche Stellungnahmeverfahren wurde am 28. September 2021 eine mündliche Anhörung durchgeführt.

In den schriftlichen und mündlichen Stellungnahmen wurden umfangreiche Einwände hinsichtlich der beabsichtigten Änderung der Verfahrensordnung eingebracht.

In den schriftlichen und mündlichen Stellungnahmen wurden von den Beteiligten Vorschläge zum weiteren Vorgehen des G-BA in Bezug auf die geplante Änderung der Modulvorlage unterbreitet. Dies waren im Wesentlichen: (1.) die geplanten Änderungen in der VerfO nicht

zu übernehmen bzw. (2.) die Anwendung des Responsekriteriums von mindestens 15 % der Skalenspannweite nur für die Fallkonstellationen vorzusehen, in denen keine etablierte/akzeptierte MID vorliegt bzw. (3.) die Fortführung der Diskussion und Entwicklung eines Kriterienkataloges zur Beurteilung der MID. Zudem wurde vorgeschlagen, alle MIDs um eine einheitliche „Sicherheitsspanne“ zu ergänzen und somit den diskutierten Unsicherheiten zu begegnen.

Diese oben genannten Vorschläge adressieren im überwiegenden Teil ein Fortführen der derzeitigen Praxis. Die Einwände der Stellungnehmenden haben insbesondere abgestellt auf die Auswirkungen der Anwendung einer Responseschwelle von 15 % auf verschiedene Fragebögen, das methodische Vorgehen zur Ableitung der Responseschwelle von 15 %, die aktuelle wissenschaftliche Diskussion zur Bewertung von MIDs sowie auf die Gegenüberstellung von Ergebnissen akzeptierter MIDs aus Beschlüssen des G-BA mit einer Responseschwelle von 15 %.

Die im Rahmen des Stellungnahmeverfahrens vorgetragenen Argumente wurden in der Zusammenfassenden Dokumentation gewürdigt.

Insgesamt bleibt festzustellen, dass die Studien zur Bestimmung von MIDs in der Regel nicht dem aktuellen Stand der wissenschaftlichen Erkenntnis entsprechen. Dementsprechend gehen Responderanalysen auf Basis eines Responsekriteriums im Sinne einer MID mit wesentlichen Unsicherheiten bei der Interpretation der klinischen Relevanz des beobachteten Effektes einher. Diese methodischen Unsicherheiten konnten von den Stellungnehmenden nicht ausgeräumt oder entkräftet werden. Zudem wurden keine substantiierten Alternativvorschläge unterbreitet, wie man eine zuverlässige Aussage zu einer klinisch relevanten Veränderung ableiten könnte.

Es zeigte sich ein allgemeiner Konsens, dass Responderanalysen allgemeine Vorteile gegenüber der Analyse stetiger Daten aufweisen.

Unbenommen der aktuellen und anhaltenden wissenschaftlichen Diskussion um Kriterien für die Entwicklung von einer MID und möglicher zukünftiger Standards ermöglicht die Anwendung eines Responsekriteriums von mindestens 15 % der Skalenspannweite zum jetzigen Zeitpunkt – im Gegensatz zu MIDs – die Berücksichtigung von aussagekräftigen Responderanalysen im Rahmen der Nutzenbewertung. Ein Wert von 15 % der Skalenspannweite soll hierbei sicherstellen, dass eine für die Patientinnen und Patienten hinreichend sicher spürbare Veränderung abgebildet wird.

In Bezug auf die Kritik am methodischen Vorgehen zur Ableitung der Responseschwelle von 15 % ist festzuhalten, dass vom IQWiG ein Wert von 15 % der Spannweite der jeweiligen Skalen als plausibler Schwellenwert für eine hinreichend sicher spürbare Veränderung auf nachvollziehbarer Grundlage hergeleitet wurde. Das Vorgehen wurde im Detail in der Anhörung und Würdigung der Stellungnahmen zum Methodenpapier 6.0 des IQWiG und in der mündlichen Anhörung zur Anpassung der Anlage II.6 zum 5. Kapitel der VerfO erörtert. U.a. erfolgte die Herleitung auf der Grundlage einer fokussierten Recherche nach systematischen Übersichtsarbeiten zu MIDs und unter Berücksichtigung der aktuellen Mindestqualitätskriterien zur Methodik der Ermittlung der MID. Es mussten u.a. folgende Kriterien erfüllt werden: longitudinale Studie, ankerbasierte MID, patientenberichteter Anker,

Global-Rating-of-Change(GRC)-Anker, der Cut-off für den GRC-Anker sollte bei minimal, small, little oder höchstens moderate liegen, um die Ermittlung einer MID zu gewährleisten. Die extrahierten MIDs wurden im Verhältnis zur Spannweite der jeweiligen Skala dargestellt (MID in % der Skalenspannweite). Die Werte am unteren Rand der erhobenen MIDs wurden vom IQWIG u.a. aufgrund der Abgrenzung von spürbarer Veränderung und Messunsicherheit als wenig geeignet eingestuft. Hinsichtlich der Frage, ob die Werte am oberen Rand des Spektrums verlässlicher sind, kommt das IQWiG zu dem Schluss, dass sich diese Frage aufgrund fehlender akzeptierter Standards zur Qualitätsbewertung, wie auch fehlender Berichtsqualität von MID-Studien, nicht beantworten lässt. Jedoch zeige die systematische Betrachtung, dass der wesentliche Anteil empirisch ermittelter MIDs unterhalb von 20 % der jeweiligen Skalenspannweite liegt. Auf Basis dieser aktuell bestverfügbaren Evidenz erfolgte vom IQWIG eine empirisch gestützte Setzung der Responseschwelle von 15 %, welche als plausibel geeignet angesehen wird, hinreichend sicher eine für Patientinnen und Patienten spürbare Veränderung abzubilden.

Die Herleitung der Responseschwelle von 15 % fand folglich unter Berücksichtigung des aktuellen methodischen Diskurses zu dieser Thematik statt, womit die Festlegung der Responseschwelle auf Basis des aktuellen Standes der wissenschaftlichen Erkenntnis vorgenommen wurde.

Hinsichtlich der Auswirkungen der Anwendung einer Responseschwelle von 15 % auf verschiedene Fragebögen ist festzustellen, dass die Definition der Responseschwelle von 15 % der Skalenspannweite dazu führt, dass zum Teil bisher verwendete Responsekriterien (MID) nicht mehr berücksichtigt werden. Dies steht der Anpassung der Anlage II.6 zum 5. Kapitel der VerfO jedoch nicht entgegen, da die Studien zur Bestimmung von MIDs in der Regel nicht dem aktuellen Stand der wissenschaftlichen Erkenntnis entsprechen. Auch kann nicht – insbesondere unter Berücksichtigung der Beispiele in der Nutzenbewertung seit Einführung des neuen methodischen Vorgehens – abgeleitet werden, dass der neu vorgeschlagene Wert von 15 % der Spannweite der jeweiligen Skalen den Nachweis von Vor- und Nachteilen von Therapien in patientenberichteten Endpunkten in relevantem Ausmaß erschwert. Es ist davon auszugehen, dass der Nachweis eines Effektes nicht von einer spezifischen MID abhängt (deren Validierung zudem aktuell in der Regel nicht dem aktuellen Stand der wissenschaftlichen Erkenntnis entspricht), sondern, dass ein Effekt auch mit dem neu vorgeschlagenen Wert von 15 % der Spannweite der jeweiligen Skalen nachweisbar ist.

Um diesen Themenkomplex im Rahmen des vorliegenden Stellungnahmeverfahrens aufzugreifen, wurden die Stellungnehmenden, die über Studiendaten verfügen, bei denen Responderanalysen im Sinne einer MID vom G-BA in abgeschlossenen Nutzenbewertungen berücksichtigt wurden, gebeten, eine Gegenüberstellung dieser Ergebnisse mit denen einer Responseschwelle von 15 % der Skalenspannweite des Instruments in das Stellungnahmeverfahren einzubringen.

Von den Stellungnehmenden wurde keine Gegenüberstellung von Ergebnissen akzeptierter MIDs aus Beschlüssen des G-BA mit einer Responseschwelle von 15 % vorgenommen. Stattdessen wurde von den Stellungnehmenden auf Simulationen abgestellt. In der Simulation zeigte sich laut den Stellungnehmenden, dass je nach Szenario für die 10 % Responseschwelle

sowohl ein Power-Gewinn als auch ein Power-Verlust verglichen mit der 15 % Responseschwelle möglich war. Jedoch lag in den meisten untersuchten Szenarien eine höhere Power für die 10 % Responseschwelle gegenüber der 15 % Responseschwelle vor. Dies galt z.B. insbesondere bei Szenarien bei schiefer Baseline-Verteilung.

Im Rahmen der Bewertung der von den Stellungnehmenden vorgelegten Simulations-Untersuchengen ist jedoch festzustellen, dass unklar bleibt, ob die Powerverschiebungen, die sich in der Simulationsstudie zeigen, Auswirkungen auf die Nutzenbewertung haben. So bleibt offen, in welchen Arealen des Parameterraums man sich in der Praxis – im Rahmen der in der Nutzenbewertung – bewegt, z.B. in wie vielen Dossiers und patientenrelevanten Endpunkten schiefe Baseline-Verteilungen vorliegen. Zur Beantwortung der Frage, welche in der Simulationsstudie identifizierten Konstellationen praxisrelevant sind, ist eine entsprechende Empirie notwendig. Zudem stehen den Simulationsszenarien mit einem Powerverlust aufgrund der 15 % Responseschwelle Szenarien gegenüber, in denen es zu einem Powergewinn kommt.

Insgesamt kann aus der Simulationsstudie daher nicht abgeleitet werden, dass der neu vorgeschlagene Wert von 15 % der Spannweite der jeweiligen Skalen den Nachweis von Vor- und Nachteilen von Therapien in patientenberichteten Endpunkten erschwert.

Abgesehen von den hier beschriebenen methodischen Unsicherheiten ist nicht nachvollziehbar, dass eine Simulation aussagekräftiger ist als ein Ansatz über konkrete Beispiele. Es wurden keine konkreten Beispiele vorlegt, die das nun durch IQWiG und G-BA vorgeschlagene Vorgehen infrage stellen. Auf der anderen Seite zeigen die Erfahrungen in den Nutzenbewertungsverfahren seit Anpassung des Methodenpapiers, dass es sich um ein praktikables Vorgehen handelt. In diesem Zusammenhang wird auch auf die fehlenden praktischen Konsequenzen eines Wechsels des Responsekriteriums von 10 Punkten auf 15 % für die EORTC-Fragebögen hingewiesen. Die von den Stellungnehmenden vorgelegte Simulation berücksichtigt den tatsächlichen Aufbau der in der Wissenschaft etablierten EORTC-Fragebögen nicht und ist daher für Aussagen, welche Auswirkungen unterschiedliche Responsekriterien auf die Ergebnisse zu diesen Fragebögen haben, ungeeignet.

Ein weiterer Vorteil des neuen Vorgehens liegt bei Instrumenten mit komplexen Skalen, für die keine bisher (nach dem alten Vorgehen) validierten MIDs identifiziert werden konnte. Für diese Instrumente kann jetzt auch – sofern angezeigt – post hoc ein Responsekriterium von genau 15 % der Skalenspannweite berechnet werden. Dies schließt eine Lücke, da für diese Instrumente im Rahmen der Nutzenbewertung bisher keine Responderanalyse berechnet werden konnte.

In der Gesamtschau der in das Stellungnahmeverfahren eingebrachten Argumente konnten keine tragfähigen Alternativen gegenüber der Anpassung der Anlage II.6 zum 5. Kapitel der VerfO identifiziert werden.

Aus der Auswertung der schriftlichen Stellungnahmen und der mündlichen Anhörung ergaben sich folglich keine Änderung der beabsichtigten Anpassung der Anlage II.6 zum 5. Kapitel der VerfO.

Gegenüber dem im Stellungnahmeverfahren gegenständlichen Beschlussentwurf wurden im Anschluss an das Stellungnahmeverfahren lediglich klarstellende Anpassungen vorgenommen. Demnach sind für den Fall, dass sowohl geeignete Responderanalysen (Responsekriterium präspezifiziert mindestens 15 % der Skalenspannweite oder post hoc genau 15 % der Skalenspannweite) als auch Analysen stetiger Daten vorliegen, die Responderanalysen darzustellen.

Eine wie von den Stellungnehmenden gewünschte Fortführung der Diskussion um die Entwicklung eines Kriterienkataloges zur Beurteilung der MID findet aktuell und auch zukünftig, z.B. im SISAQOL-Projekt (Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data-Consortium), statt.

Wenn ein neues methodisch hochwertiges Vorgehen zur Validierung von klinischen Responseschwellen etabliert werden konnte, wird der G-BA die geforderten methodischen Anforderungen überprüfen und anpassen.

3 Bürokratiekostenermittlung

Durch den vorgesehenen Beschluss entstehen keine neuen bzw. geänderten Informationspflichten für Leistungserbringer im Sinne von Anlage II zum 1. Kapitel VerfO und dementsprechend keine Bürokratiekosten.

4 Verfahrensablauf

Der Unterausschuss Arzneimittel hat zur Vorbereitung einer Überarbeitung der VerfO zur Änderung der Anlage II.6 zum 5. Kapitel und Erstellung einer Beschlussempfehlung zur Einleitung eines Stellungnahmeverfahrens zur Änderung der VerfO im 5. Kapitel die Arbeitsgruppe Entscheidungsgrundlagen beauftragt.

Der Unterausschuss Arzneimittel hat in seiner Sitzung am 8. Juni 2021 über die Änderungen im 5. Kapitel der VerfO beraten und die Beschlussvorlage über die Einleitung eines Stellungnahmeverfahrens zur Änderung der VerfO im 5. Kapitel konsentiert.

Die Beschlussvorlage wurde der Arbeitsgruppe Geschäftsordnung-Verfahrensordnung übersandt, die am 10. Juni 2021 schriftlich über die Beschlussunterlagen abgestimmt und diese an das Plenum des Gemeinsamen Bundesausschusses zur Beschlussfassung nach 1. Kapitel § 8 Absatz 2 Satz 1 Buchstabe b VerfO weitergeleitet hat.

Das Plenum des Gemeinsamen Bundesausschusses hat am 17. Juni 2021 über die Beschlussempfehlungen zur Einleitung eines Stellungnahmeverfahrens zur Änderung der VerfO im 5. Kapitel beraten und die Einleitung eines Stellungnahmeverfahrens zur Änderung der VerfO im 5. Kapitel zur Änderung der Modulvorlage in der Anlage II beschlossen.

Die mündliche Anhörung wurde am 28. September 2021 durchgeführt.

Die Beschlussvorlage zur Änderung der VerfO im 5. Kapitel wurde in der Sitzung des Unterausschusses Arzneimittel am 7. Dezember 2021 konsentiert und der Arbeitsgruppe Geschäftsordnung-Verfahrensordnung übersandt.

Die Arbeitsgruppe Geschäftsordnung-Verfahrensordnung hat in ihrer Sitzung am 10. Dezember 2021 über die Beschlussunterlagen abgestimmt und diese an das Plenum des Gemeinsamen Bundesausschusses zur Beschlussfassung weitergeleitet.

Das Plenum des Gemeinsamen Bundesausschusses hat am 16. Dezember 2021 die Änderung der VerfO im 5. Kapitel beschlossen.

Zeitlicher Beratungsverlauf

Sitzung	**Datum**	**Beratungsgegenstand**
AG Entscheidungsgrundlagen	3. Mai 2021 31. Mai 2021	Beratung über die Änderungen der Anlage II.6 zum 5. Kapitel der Verfahrensordnung und Erstellung einer Beschlussempfehlung zur Einleitung eines diesbezüglichen Stellungnahmeverfahrens
Unterausschuss Arzneimittel	8. Juni 2021	Beratung und Konsentierung der Beschlussvorlage zur Einleitung des Stellungnahmeverfahrens hinsichtlich der Änderung der Anlage II.6 zum 5. Kapitel der Verfahrensordnung
AG Geschäftsordnung-Verfahrensordnung	10. Juni 2021	Schriftliche Abstimmung über die Beschlussvorlage
Plenum	17. Juni 2021	Beschlussfassung zur Einleitung des Stellungnahmeverfahrens hinsichtlich der Änderung der Anlage II.6 zum 5. Kapitel der Verfahrensordnung
Unterausschuss Arzneimittel	28. September 2021	Durchführung der mündlichen Anhörung
AG Entscheidungsgrundlagen	14. Oktober 2021 11. November 2021	Auswertung des Stellungnahmeverfahrens und Beratung zur Beschlussfassung hinsichtlich der Änderung der Anlage II.6 zum 5. Kapitel der Verfahrensordnung
Unterausschuss Arzneimittel	7. Dezember 2021	Beratung und Konsentierung der Beschlussvorlage
AG Geschäftsordnung-Verfahrensordnung	10. Dezember 2021	Abstimmung über die Beschlussvorlage
Plenum	16. Dezember 2021	Beschlussfassung hinsichtlich der Änderung der Anlage II.6 zum 5. Kapitel der Verfahrensordnung

Berlin, den 16. Dezember 2021

Gemeinsamer Bundesausschuss
gemäß § 91 SGB V
Der Vorsitzende

Prof. Hecken

5 Beschluss

Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Verfahrensordnung:

Änderung der Modulvorlage in der Anlage II zum 5. Kapitel

Vom 16. Dezember 2021

Der Gemeinsame Bundesausschusses hat in seiner Sitzung am 16. Dezember 2021 beschlossen, die Anlage II zum 5. Kapitel der Verfahrensordnung in der Fassung vom 18. Dezember 2008 (BAnz. Nr. 84a vom 10. Juni 2009), die zuletzt durch die Bekanntmachung des Beschlusses vom T. Monat JJJJ (BAnz AT TT.MM.JJJJ BX) geändert worden ist, wie folgt zu ändern:

I. In Anlage II.6 (Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen) werden in Unterabschnitt 4.3.1.3.1 (<Endpunkt xxx> – RCT) nach dem Satz „Zu mit Skalen erhobenen patientenberichteten Endpunkten (z.B. zur gesundheitsbezogenen Lebensqualität oder zu Symptomen) sind immer auch die Werte im Studienverlauf anzugeben, auch als grafische Darstellung, sowie eine Auswertung, die die über den Studienverlauf ermittelten Informationen vollständig berücksichtigt (z.B. als Symptomlast über die Zeit, geschätzt mittels MMRM-Analyse [falls aufgrund der Datenlage geeignet]).“ folgende Sätze eingefügt:

„Die Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen soll nach dem folgenden Vorgehen erfolgen:

1. Falls in einer Studie Responderanalysen unter Verwendung einer MID präspezifiziert sind und das Responsekriterium mindestens 15 % der Skalenspannweite des verwendeten Erhebungsinstruments entspricht, sind diese Responderanalysen für die Bewertung darzustellen.

2. Falls präspezifiziert Responsekriterien im Sinne einer MID unterhalb von 15 % der Skalenspannweite liegen, bestehen in diesen Fällen und solchen, in denen gar keine Responsekriterien präspezifiziert wurden, aber stattdessen Analysen kontinuierlicher Daten zur Verfügung stehen, verschiedene Möglichkeiten. Entweder können post hoc spezifizierte Analysen mit einem Responsekriterium von genau 15 % der Skalenspannweite dargestellt werden. Alternativ können Analysen der kontinuierlichen Daten dargestellt werden, für die Relevanzbewertung ist dabei auf ein allgemeines statistisches Maß in Form von standardisierten Mittelwertdifferenzen (SMDs, in Form von Hedges’ g) zurückzugreifen. Dabei ist eine Irrelevanzschwelle als Intervall von - 0,2 bis 0,2 zu verwenden: Liegt das zum Effektschätzer korrespondierende Konfidenzintervall vollständig außerhalb dieses Irrelevanzbereichs, wird davon ausgegangen, dass die Effektstärke nicht in einem sicher irrelevanten Bereich liegt. Dies soll gewährleisten, dass der Effekt hinreichend sicher mindestens als klein angesehen werden kann.

3. Liegen sowohl geeignete Responderanalysen (Responsekriterium präspezifiziert mindestens 15 % der Skalenspannweite oder post hoc genau 15 % der Skalenspannweite) als auch Analysen stetiger Daten vor, sind die Responderanalysen darzustellen."

II. Die Änderung der Verfahrensordnung tritt am Tag nach der Veröffentlichung im Bundesanzeiger in Kraft.

Die Tragenden Gründe zu diesem Beschluss werden auf den Internetseiten des Gemeinsamen Bundesausschusses unter www.g-ba.de veröffentlicht.

Berlin, den 16. Dezember 2021

Gemeinsamer Bundesausschuss
gemäß § 91 SGB V
Der Vorsitzende

Prof. Hecken

B. Dokumentation des Stellungnahmeverfahrens

Gemäß 1. Kapitel § 8 Absatz 2 Satz 1 Buchstabe b der Verfahrensordnung (VerfO) kann das Plenum im Einzelfall beschließen, dass zu Entscheidungen, bei denen kein gesetzlich eingeräumtes Stellungnahmerecht besteht, ebenfalls Stellungnahmen einzuholen sind.

Der G-BA hat in seiner Sitzung am 17. Juni 2021 beschlossen, ein Stellungnahmeverfahren zur Änderung der Modulvorlagen in der Anlage II zum 5. Kapitel der VerfO einzuleiten.

Bezüglich Änderungen der Arzneimittel-Richtlinie aufgrund von Beschlüssen zur frühen Nutzenbewertung von Arzneimitteln sind die Sachverständigen der medizinischen und pharmazeutischen Wissenschaft und Praxis sowie die für die Wahrnehmung der wirtschaftlichen Interessen gebildeten maßgeblichen Spitzenorganisationen der pharmazeutischen Unternehmer, die betroffenen pharmazeutischen Unternehmer, die Berufsvertretungen der Apotheker und die maßgeblichen Dachverbände der Ärztegesellschaften der besonderen Therapierichtungen auf Bundesebene gemäß § 35a Absatz 3 Satz 2 i.V.m. § 92 Absatz 3a SGB V stellungnahmeberechtigt. Unter entsprechender Anwendung dieser Stellungnahmerechte wurde der Beschlussentwurf zur Änderung der Anlage II.6 zum 5. Kapitel der VerfO den folgenden Organisationen sowie den Verbänden der pharmazeutischen Unternehmen mit der Bitte um Weiterleitung zugesendet.

Folgende Organisationen wurden angeschrieben:

Organisation	**Straße**	**Ort**
Bundesverband der Pharmazeutischen Industrie e. V. (BPI)	Friedrichstr. 148	10117 Berlin
Verband Forschender Arzneimittelhersteller e. V. (vfa)	Hausvogteiplatz 13	10117 Berlin
Bundesverband der Arzneimittel-Importeure e. V. (BAI)	EurimPark 8	83416 Saaldorf-Surheim
Bundesverband der Arzneimittel-Hersteller e. V. (BAH)	Friedrichstr. 134	10117 Berlin
Biotechnologie-Industrie-Organisation Deutschland e. V. (BIO Deutschland e. V.)	Schützenstraße 6a	10117 Berlin
Pro Generika e. V.	Unter den Linden 32 - 34	10117 Berlin
Arzneimittelkommission der Deutschen Ärzteschaft (AkdÄ)	Herbert-Lewin-Platz 1	10623 Berlin
Arzneimittelkommission der Deutschen Zahnärzteschaft (AK-Z) c/o Bundeszahnärztekammer	Chausseestr. 13	10115 Berlin
Bundesvereinigung Deutscher Apothekerverbände e. V. (ABDA)	Heidestr. 7	10557 Berlin
Deutscher Zentralverein Homöopathischer Ärzte e. V.	Axel-Springer-Str. 54b	10117 Berlin

Gesellschaft Anthroposophischer Ärzte e. V.	Herzog-Heinrich-Str. 18	80336 München
Gesellschaft für Phytotherapie e. V.	Postfach 10 08 88	18055 Rostock

Darüber hinaus wurde die Einleitung des Stellungnahmeverfahrens im Bundesanzeiger bekanntgemacht (BAnz AT 25.06.2021 B5).

(siehe C. Anhang der Zusammenfassenden Dokumentation)

1 Unterlagen des Stellungnahmeverfahrens

1.1 Schriftliches Stellungnahmeverfahren

(siehe C. Anhang der Zusammenfassenden Dokumentation)

1.2 Mündliche Anhörung

Mit Datum vom 6. September 2021 wurden die pharmazeutischen Unternehmer/Organisationen, die berechtigt sind, zu einem Beschluss des Gemeinsamen Bundesausschusses Stellung zu nehmen und eine schriftliche Stellungnahme abgegeben sowie ihre Teilnahme mit der schriftlichen Stellungnahme angemeldet haben, zu der mündlichen Anhörung eingeladen.

2 Übersicht der eingegangenen Stellungnahmen

2.1 Übersicht der eingegangenen schriftlichen Stellungnahmen

Organisation	Eingangsdatum
UCB Pharma GmbH	09.07.2021
AMGEN GmbH	13.07.2021
Novo Nordisk Pharma GmbH	13.07.2021
Merck Serono GmbH	15.07.2021
Bundesverband der Pharmazeutischen Industrie e. V.	15.07.2021
Deutsche Diabetes Gesellschaft e. V. (DDG)	20.07.2021
Deutsche Atemwegsliga e. V.	21.07.2021
Deutsche Gesellschaft für Innere Medizin e. V. (DGIM)	21.07.2021
Collegium Internationale Psychiatriae Scalarum (CIPS), Dr. Lorkowski	21.07.2021
Bristol-Myers Squibb GmbH & Co. KGaA	21.07.2021
AbbVie Deutschland GmbH & Co. KG	22.07.2021

Roche Pharma AG	22.07.2021
Novartis Pharma GmbH	22.07.2021
Verband Forschender Arzneimittelhersteller e. V. (vfa)	22.07.2021
Astellas Pharma GmbH	22.07.2021
IQVIA Commercial GmbH & Co. OHG	22.07.2021
Ecker + Ecker GmbH	23.07.2021
Deutsche Gesellschaft für Psychiatrie und Psychotherapie, Psychosomatik und Nervenheilkunde e. V. (DGPPN)	23.07.2021
Pfizer Deutschland GmbH	23.07.2021
Bayer Vital GmbH	23.07.2021
GlaxoSmithKline GmbH & Co. KG	23.07.2021
MSD Sharp & Dohme GmbH	23.07.2021
Deutsche Gesellschaft für Psychologische Schmerztherapie und -forschung e. V. (DGPSF)	23.07.2021
Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF)	23.07.2021
Janssen-Cilag GmbH	23.07.2021
Deutsche Schmerzgesellschaft e. V.	23.07.2021
Biogen GmbH	23.07.2021
Boehringer Ingelheim Pharma GmbH & Co. KG	24.07.2021

2.2 Übersicht der Anmeldung zur mündlichen Anhörung

Organisation	Name
UCB Pharma GmbH	Andreas, Hr.
AMGEN GmbH	Stein, Fr. Floßmann, Fr. Dr.
Novo Nordisk Pharma GmbH	Bauer, Hr. Dr. Kiencke, Hr. Dr.
Merck Serono GmbH	Schlichting, Hr. Osowski, Fr. Dr.
Bundesverband der Pharmazeutischen Industrie e. V.	Wilken, Hr. Dr.
Deutsche Atemwegsliga e. V.	Kardos, Hr. Dr. Worth, Prof. Dr.
Deutsche Diabetes Gesellschaft e. V. (DDG)	Gallwitz, Hr. Prof. Dr. Müller-Wieland, Hr. Prof. Dr.
Deutsche Gesellschaft für Innere Medizin e. V. (DGIM)	Sauerbruch, Hr. Prof. Dr.
Collegium Internationale Psychiatriae Scalarum (CIPS), Dr. Lorkowski	Lorkowski, Hr. Dr.

Bristol-Myers Squibb GmbH & Co. KGaA	Kupas, Fr. Dr.
AbbVie Deutschland GmbH & Co. KG	Sternberg, Fr. Dr. Gossens, Hr.
Roche Pharma AG	Csintalan, Hr. Dr. Knoerzer, Hr. Dr.
Novartis Pharma GmbH	Marx, Fr. Dr. Eichinger, Fr. Dr.
Verband Forschender Arzneimittelhersteller e. V. (vfa)	Rasch. Hr. Dr.
Astellas Pharma GmbH	Zölch, Fr.
IQVIA Commercial GmbH & Co. OHG	Böhm, Fr.
Pfizer Deutschland GmbH	Miller, Hr. Dr. Leverkus, Hr.
Bayer Vital GmbH	Dintsios, Hr. Dr.
GlaxoSmithKline GmbH & Co. KG	Hennig, Hr. PD Dr. Karl, Hr. Dr.
MSD Sharp & Dohme GmbH	Rettelbach, Fr. Ziegler, Hr. Dr.
Janssen-Cilag GmbH	Huschens, Fr. Dr.
Deutsche Schmerzgesellschaft e. V.	Isenberg, Hr.
Biogen GmbH	Plesnila-Frank, Fr. Dichter, Hr. Dr.
Boehringer Ingelheim Pharma GmbH & Co. KG	Pfarr, Hr. Henschel, Hr. Dr.

2.2.1 Zusammenfassende Angaben der Offenlegungserklärung

Organisation, Name	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6
UCB Pharma GmbH						
Andreas, Hr.	ja	nein	nein	nein	nein	ja
Amgen GmbH						
Stein, Fr.	ja	nein	nein	nein	nein	ja
Floßmann, Fr. Dr.	ja	nein	nein	nein	nein	ja
Novo Nordisk Pharma GmbH						
Bauer, Hr. Dr.	ja	nein	nein	nein	nein	ja
Kiencke, Hr. Dr.	ja	nein	nein	nein	nein	nein
Bundesverband der Pharmazeutischen Industrie e. V.						
Wilken, Hr. Dr.	ja	nein	nein	nein	nein	nein
Deutsche Atemwegsliga e. V.						
Kardos, Hr. Dr.	nein	ja	ja	nein	nein	nein

Organisation, Name	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6
Worth, Hr. Prof. Dr	nein	ja	ja	nein	nein	nein
Deutsche Gesellschaft für Innere Medizin e. V. (DGIM)						
Sauerbruch, Hr. Prof. Dr.	ja	ja	ja	nein	ja	nein
Collegium Internationale Psychiatriae Scalarum (CIPS), Dr. Lorkowski						
Weyer, Hr. Prof. Dr.	ja	ja	nein	nein	nein	ja
Görtelmeyer, Hr. Prof. Dr.	nein	ja	ja	nein	ja	ja
Bristol-Myers Squibb GmbH & Co. KGaA						
Kupas, Fr. Dr.	ja	nein	nein	nein	nein	ja
AbbVie Deutschland GmbH & Co. KG						
Sternberg, Fr. Dr.	ja	nein	nein	nein	nein	nein
Gossens, Hr.	ja	nein	nein	nein	nein	nein
Roche Pharma AG						
Csintalan, Hr. Dr.	ja	nein	nein	nein	nein	ja
Knoerzer, Hr. Dr.	ja	nein	nein	nein	nein	ja
Novartis Pharma GmbH						
Marx, Fr. Dr.	ja	ja	nein	nein	nein	nein
Eichinger, Fr. Dr.	ja	nein	nein	nein	nein	ja
Astellas Pharma GmbH						
Zölch, Fr.	ja	nein	nein	nein	nein	nein
Groß-Langenhoff, Hr. Dr.	ja	nein	nein	nein	nein	nein
IQVIA Commercial GmbH & Co. OHG						
Böhm, Fr.	nein	ja	nein	nein	nein	nein
Pfizer Deutschland GmbH						
Miller, Hr. Dr.	ja	nein	nein	nein	nein	nein
Leverkus, Hr.	ja	nein	nein	nein	nein	ja
Bayer Vital GmbH						
Dintsios, Hr. Dr.	ja	nein	nein	nein	nein	nein
GlaxoSmithKline GmbH & Co. KG						
Hennig, Hr. PD Dr.	ja	nein	nein	nein	nein	ja
Karl, Hr. Dr.	ja	nein	nein	nein	nein	nein
MSD Sharp & Dohme GmbH						
Rettelbach, Fr.	ja	nein	nein	nein	nein	ja
Ziegler, Hr. Dr.	ja	nein	nein	nein	nein	ja

Organisation, Name	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6
Verband Forschender Arzneimittelhersteller e. V. (vfa)						
Rasch, Hr. Dr.	ja	nein	nein	nein	nein	nein
Janssen-Cilag GmbH						
Huschens, Fr. Dr.	ja	nein	nein	nein	nein	ja
Biogen GmbH						
Plesnila-Frank, Fr.	ja	nein	nein	nein	nein	ja
Dichter, Hr. Dr.	ja	nein	nein	nein	nein	ja
Boehringer Ingelheim Pharma GmbH & Co. KG						
Pfarr, Hr.	ja	ja	ja	nein	nein	nein
Henschel, Hr. Dr.	ja	nein	nein	nein	nein	nein
Boehringer Ingelheim Pharma GmbH & Co. KG						
Pfarr, Hr.	ja	ja	ja	nein	nein	nein
Henschel, Hr. Dr.	ja	nein	nein	nein	nein	nein
Deutsche Diabetes Gesellschaft e. V. (DDG)						
Gallwitz, Hr. Prof. Dr.	nein	ja	ja	nein	nein	nein
Müller-Wieland, Hr. Prof. Dr.	nein	ja	ja	ja	ja	nein
Merck Serono GmbH						
Schlichting, Hr.	ja	nein	nein	nein	nein	ja
Osowski, Fr. Dr.	ja	nein	nein	nein	nein	ja

2.3 Auswertung der Stellungnahmen

2.3.1 Auswirkungen der Anwendung einer Responseschwelle von 15 % auf verschiedene Fragebögen

Einwand

Saint-George's Respiratory Questionnaire (SGRQ)

GlaxoSmithKline GmbH & Co. KG

Für das in der Indikation der chronisch obstruktiven Lungenerkrankung (COPD) sehr häufig verwendete Erhebungsinstrument Saint-George's Respiratory Questionnaire (SGRQ) würde die Anwendung der 15 % Schwelle zu einem Schwellenwert von 15 führen, da die Skalenspannweite 0-100 beträgt. Die wissenschaftlich etablierte und in bisherigen Nutzenbewertungsverfahren akzeptierte MID liegt jedoch bei einem Wert von 4[1]. Somit würde durch die 15 % Schwelle eine Erhöhung um den Faktor 3,75 resultieren (von bisher 4 auf nunmehr 15). Entsprechende Responderanalysen auf Basis der etablierten MID wurden jedoch kürzlich vom IQWiG nicht mehr herangezogen, da die MID vom IQWiG als nicht hinreichend validiert bewertet wird[2]. Diese Vorgehensweise entbehrt einer wissenschaftlichen und evidenzbasierten Grundlage.

Der Entwickler des SGRQ, Professor Paul Jones, hat sich im Auftrag von GSK in einer separaten Stellungnahme zum wissenschaftlichen Stellenwert des IQWiG-Vorschlags geäußert[3]. Diese Stellungnahme fügen wir als Anlage bei. Professor Jones kommt dabei zu folgenden Einschätzungen:

- Es gibt eindeutige und konsistente Evidenz dafür, dass beim SGRQ eine Veränderung um 4 Einheiten einem Unterschied entspricht, der für Patienten und Ärzte von Bedeutung ist.
- Klinische Kriterien-Studien zeigen, dass beim SGRQ eine Veränderung von 4 Einheiten einhergeht mit dem Unterschied, ob der Patient ans Haus gebunden ist oder nicht oder mit dem Risiko eines bedeutsamen Ereignisses wie die Wiederaufnahme ins Krankenhaus oder den Tod
- Einige Studien geben für den SGRQ höhere Werte als 4 Einheiten für den MID an, aber selbst diese Schätzungen betragen nur die Hälfte der vom IQWiG vorgeschlagenen 15%.

Deutschen Atemwegsliga e. V.

Diese Stellungnahme bezieht sich auf die MCID- Werte (minimal clinical important difference) bei komplexen PRO („patient reported outcome")-Skalen, wie dem SGRQ (St. George's Respiratory Questionnaire) oder dem LCQ (Leicester Cough Questionnaire) und ähnlichen Instrumente.

„Gemäß dem aktuellen methodischen Vorgehen des IQWiG (Methodenpapier 6.0, veröffentlicht am 05.11.2020) erachtet das IQWiG für patientenberichtete Endpunkte eine Responseschwelle für Responder-analysen von mindestens 15 % der Skalenspannweite eines

1 Jones PW. St. George's respiratory questionnaire: MCID. COPD: Journal of Chronic Obstructive Pulmonary Disease. 2005; 2(1): 75-9.

2 G-BA, Gemeinsamer Bundesausschuss. Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII – Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V Nintedanib (interstitielle Lungenerkrankung mit systemischer Sklerose (SSc-ILD)) 2021 15.07.2021. Available from: https://www.g-ba.de/downloads/40-268-7287/2021-02-04_AM-RL-XII_Nintedanib-PF-ILDs_D-568_TrG.pdf.

3 Jones PW. The SGRQ MID. 2021 22.07.2021.

Instrumentes (bei post hoc durchgeführten Analysen von genau 15 % der Skalenspannweite) als notwendig, um eine für Patienten spürbare Veränderung hinreichend sicher abzubilden".

Begründet wird dieses Vorgehen damit, „dass zu einzelnen Instrumenten häufig eine Vielzahl von MCIDs publiziert werden, die innerhalb eines Erhebungsinstruments große Spannweiten haben können."

Die zitierte Literatur für die Begründung dieses Vorgehens bezieht sich auf ein Cochrane Database Review von 3000 MCID's sowie auf Daten zu degenerativen Gelenkerkrankungen und Fatigue-Syndrom. Die einzige Referenz auf dem Gebiet der Atemwegs- und Lungenkrankheiten ist die Arbeit von Alma et al [4], die aus klinischer Sicht retrospektive explorative Subgruppenanalysen für verschiedene Settings (ambulant, Rehabilitation) und für verschiedene Gruppen von COPD Patienten durchführte. Die Arbeit zeigt im Wesentlichen, dass für den häufig zur Erfassung der Lebensqualität benutzten Fragebogen SGRQ nicht selbstverständlich Linearität der MCID-Daten angenommen werden kann. Die MCID ist beispielsweise bei schwer erkrankten Patienten größer als die ursprünglich publizierten 4 Einheiten [5]. Die mediane Änderung des SGRQ, die gewöhnlich bei COPD- Studienpatienten bei Auswertung durch IQWIG in Betracht gezogen wird, betrug bei ambulanten Patienten −3.04 (−5.52 bis−0.57). Nur bei Reha Patienten lag sie bei −8.71 (−9.79 bis −7.63).

In der Untersuchung von Alma wurde in keinem Fall die 15% IQWIG Score Änderung (15 Punkte auf einer Skala von 0-100) erreicht, nicht einmal in Extremfällen.

Eine MCID für den SGRQ von 15 Punkten fand sich unseres Wissens bislang in keiner der von IQWIG analysierten pneumologischen Studien. Unter pharmakologischen Therapien haben Biologika bei schwerem Asthma die höchsten MCID's bezüglich des SGRQ erreicht, sie lag etwa bei 8.

Aus klinischer Sicht ist es nicht nachvollziehbar, warum eine pauschale Festlegung einer einheitlichen MCID, die aus verschiedenen Studien aus verschiedenen Fachgebieten summarisch abgeleitet wird, zu genaueren Ergebnissen führen sollte als eine individuelle Betrachtung der betroffenen Patientengruppen. Die Patienten kommen bei diesem Verfahren gar nicht erst zur Sprache, obwohl es sich um ein „patient reported outcome" handelt, das in allererster Linie die subjektive Meinung des Patienten abbilden sollte. Die Einschätzung des behandelnden Arztes, der in direktem Kontakt zum Patienten steht, sollte ebenfalls Berücksichtigung finden. Der praktische Bezug auf für Patienten wichtige Änderungen der Lebensqualität (z.B. der Patient ist unter der Therapie X jetzt in der Lage, aus dem Haus zu gehen, um kleinere Einkäufe zu tätigen oder ob das Abhusten von Sekret leichter fällt) fehlt bei der rein statistischen Berechnung der MCID ebenfalls.

Gegen die summarische Festlegung einer MCID Schwelle spricht auch, dass selbst bei einer einzigen Erkrankung wie der COPD es wesentliche Inhomogenitäten bei den zu beurteilenden Patientengruppen gibt, worauf das IQWIG korrekt hingewiesen und die ALMA Studie[4] hierzu als Referenz aufgeführt hat.

Paul Jones weist darauf hin[5], dass sowohl die Anchor-basierte als auch die Distributions-basierte Methode zur Festlegung der MCID zwangsläufig sowohl „sampling errors" als auch „measurment errors" und subjektive Beurteilungen zumindest der zugrunde gelegten Variablen beinhaltet, weshalb auf die nachfolgende Studienlage bei der Bewährung einer MCID zusätzlich geachtet werden müsse. Die klinischen Studien werden in der Regel bei stabilen ambulanten COPD Patienten durchgeführt, die sich sowohl in Hinblick auf den Schweregrad der Erkrankung (SGRQ Score um 40) als auch bezüglich des Alters den Patienten ähneln, bei denen die MCID Studie von Jones ursprünglich durchgeführt wurde. Hier hat sich in zahlreichen Studien die MCID- Schwelle von mindestens 4 Einheiten als weit akzeptiertes

4 Alma H, de Jong C, Jelusic D et al. Baseline health status and setting impacted minimal clinically important differences in COPD: an exploratory study. J Clin Epidemiol 2019; 116: 49-61

5 J Jones PW. St. George's Respiratory Questionnaire: MCID. COPD 2005; 2

Resultat etabliert. So kann eine Differenz von mindestens 4 Einheiten im SGRQ unterscheiden, ob ein Patient infolge der Schwere seiner COPD an seine Wohnung gebunden ist oder nicht oder ob ein erhöhtes Risiko für eine stationäre Aufnahme oder sogar das Eintreten des Todesfalles besteht oder nicht [6] Für abweichende Diagnosen (zum Beispiel idiopathische Lungenfibrose oder sehr schwere stationär behandelte COPD Patienten, Reha-Patienten etc.) müssen in der Zukunft angepasste MCID Scores festgelegt werden.

Bewertung

Systematische Zusammenstellungen empirisch ermittelter MIDs zeigen, dass zu einzelnen Instrumenten häufig eine Vielzahl von MIDs publiziert werden, die innerhalb eines Erhebungsinstruments große Spannweiten haben können[7, 8, 9, 10]. Ursächlich hierfür können unter anderem die in den Studien eingesetzten unterschiedlichen Anker, Beobachtungsperioden oder analytische Methoden sein[10, 11, 12].

Es existiert derzeit kein etablierter Standard, mit dem die Qualität von Studien zur Ermittlung einer MID bewertet und die Aussagekraft der ermittelten MIDs abgeschätzt werden kann[7, 11, 15]. Gleichwohl liegt ein erster Vorschlag für ein entsprechendes Instrument vor[13]. Unbenommen dessen ist jedoch festzustellen, dass wesentliche Anteile der Methodik bzw. der Kriterien für eine Qualitätsbewertung von Studien zur Ermittlung einer MID zumeist gar nicht berichtet werden[13]. Eine anhand methodischer Qualitätskriterien begründete Auswahl empirisch ermittelter MIDs für die Nutzenbewertung ist somit derzeit nicht zu treffen[11, 14, 15].

Neben den methodischen Faktoren beruht ein anderer Teil der Variabilität von MIDs – wie auch von den Stellungnehmern angemerkt wird – auf ihrer Abhängigkeit von Charakteristika der Patientenpopulation, in der das Instrument eingesetzt wird, sowie weiteren Kontextfaktoren. So können der Schweregrad der Erkrankung, die Art der eingesetzten Intervention oder die Frage, ob die Patientinnen und Patienten eine Verbesserung oder Verschlechterung ihrer Erkrankung erfahren, Einfluss auf die MID haben[16]. Der Umgang mit diesem Teil der Variabilität von MIDs ist ungeklärt.

---

6 Jones PW. Interpreting thresholds for a clinically significant change in health status in asthma and COPD. Eur Respir J 2002; 19.

7 Carrasco-Labra A, Devji T, Qasim A, Phillips M, Devasenapathy N, Zeraatkar D et al. Interpretation of patient-reported outcome measures: an inventory of over 3000 minimally important difference estimates and an assessment of their credibility. Cochrane Database Syst Rev 2018; (9 Suppl 1): 135-136.

8 Çelik D, Çoban Ö, Kılıçoğlu Ö. Minimal clinically important difference of commonly used hip-, knee-, foot-, and ankle-specific questionnaires: a systematic review. J Clin Epidemiol 2019; 113: 44-57.

9 Hao Q, Devji T, Zeraatkar D, Wang Y, Qasim A, Siemieniuk RAC et al. Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation. BMJ Open 2019; 9(2): e028777.

10 Nordin A, Taft C, Lundgren-Nilsson A, Dencker A. Minimal important differences for fatigue patient reported outcome measures: a systematic review. BMC Med Res Methodol 2016; 16: 62.

11 Devji T, Guyatt GH, Lytvyn L, Brignardello-Petersen R, Foroutan F, Sadeghirad B et al. Application of minimal important differences in degenerative knee disease outcomes: a systematic review and case study to inform BMJ Rapid Recommendations. BMJ Open 2017; 7(5): e015587.

12 Ousmen A, Touraine C, Deliu N, Cottone F, Bonnetain F, Efficace F et al. Distribution- and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review. Health Qual Life Outcomes 2018; 16(1): 228.

13 Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ 2020; 369: m1714.

14 Devji T, Carrasco-Labra A, Lytvyn L, Johnston B, Ebrahim S, Furukawa T et al. A new tool to measure credibility of studies determining minimally important difference estimates. Cochrane Database Syst Rev 2017; (9 Suppl 1): 58.

15 Johnston BC, Ebrahim S, Carrasco-Labra A, Furukawa TA, Patrick DL, Crawford MW et al. Minimally important difference estimates and methods: a protocol. BMJ Open 2015; 5(10): e007953.

16 Alma H, De Jong C, Jelusic D, Wittmann M, Schuler M, Kollen B et al. Baseline health status and setting impacted minimal clinically important differences in COPD: an exploratory study. J Clin Epidemiol 2019; 116: 49-61.

Die aktuell verwendeten MIDs bilden diese Variabilität jedoch nicht ab, da die entsprechenden Studien zur Ermittlung der MID diese Ansprüche nicht erfüllen.

Insgesamt gehen die genannten Limitationen bei Responderanalysen auf Basis eines Responsekriteriums im Sinne einer MID mit wesentlichen Unsicherheiten bei der Interpretation der klinischen Relevanz des beobachteten Effektes einher.

Die Anpassung der Anlage II.6 zum 5. Kapitel der VerfO stellt u.a. sicher, dass für Responderanalysen im Rahmen der Nutzenbewertung geeignete Responseschwellen eingesetzt werden, die eine für die Patientinnen und Patienten hinreichend sicher spürbare Veränderungen abbilden, womit diesbezügliche Unsicherheiten bei der Interpretation der klinischen Relevanz des beobachteten Effektes verhindert werden sollen.

Einwand

Short Form-12 Health Survey, Short Form-36 Health Survey (SF-12/SF-36)

GlaxoSmithKline GmbH & Co. KG

GSK verwendet in seinen Studien unter anderem auch das Instrument „Short Form-12 Health Survey“ (SF-12) zur Beurteilung der gesundheitsbezogenen Lebensqualität. Vor diesem Hintergrund haben wir uns an den Entwickler dieses Instrumentes, Quality Metric, mit der Frage gewendet, wie die vom IQWiG vorgeschlagene 15%-Schwelle beim SF-12 konkret implementiert werden sollte. Die entsprechende Einschätzung von Quality Metric bezieht sich sowohl auf den SF-36 als auch auf den SF-12 und wird als Anlage zu dieser Stellungnahme angefügt[17]. Hierin kommt Quality Metric zu folgendem Schluss:

“While the 15% principle seems like a good rule of thumb for many scales, we think – for the reasons discussed above - that there are measures (such as SF-36v2®PCS and MCS scores) for which this principle should not be applied. Imagine a patient that improved his score on the SF-36v2®PF, RP, BP, and GH scales by 6.1, 6.0, 6.5, and 7.3 points respectively. This is slightly more than 15% of the range for each scale. If scores on the 4 other scale were unchanged, the improvement on SF-36v2®PCS would be 8.6 points. This is a numerically larger improvement on the PCS than on any subscale. Nevertheless, the improvement on PCS would not fulfill the 15% principle, although the improvement on each of the 4 physical subscales does fulfill the 15% principle. This would not provide consistent interpretation of results.

For these reasons, we believe that it would be inappropriate to apply to the SF-36v2®PCS and MCS the principle that the response criteria should be defined by 15% of the score range. Our suggested response criteria are 3.4 and 4.6 points for SF-36v2®PCS and MCS. More conservative suggestions, based on the reliable change index, would be 5.5 points for SF-36v2® PCS and 7.5 points for MCS. For the SF-12v2®, our suggested response criteria are 6.0 for PCS and 7.0 points MCS. While the thresholds are smaller than the thresholds derived from the 15% principle (8.7 for PCS and 9.9 for MCS), the differences are smaller than for the SF-36v2®.”

Novo Nordisk Pharma GmbH

Novo Nordisk hat sich in Bezug auf das 15 %-Responsekriterium mit dem Entwicklern des generischen Fragebogens SF-36v2® ausgetauscht und teilt die Ansicht der Entwickler des SF-36v2®, dass ein für alle Fragebögen einheitlicher Schwellenwert als Responsekriterium nicht angemessen ist und insbesondere für die beiden normbasierten Summen-Scores des SF-36v2® (physical component summary (PCS) und mental component summary (MCS)) nicht angewendet werden sollte. Eine detaillierte Begründung in Hinblick auf PCS und MCS ist dem

17 Quality Metric. Quality Metric Statement. 2021 15.07.2021.

zitierten Schreiben des Chief Science Officer des SF-36v2®-Entwicklers (Jakob Bjorner, Quality Metric) zu entnehmen (Bjorner 2021, 18). […]

Ein wesentlicher Grund warum der 15 %-Schwellenwert für PCS und MCS nicht angewendet werden sollte, ist laut Bjorner 2021 der Tatsache geschuldet, dass diese beiden normbasierten Summen-Scores PCS und MCS die Informationen aller acht Subskalen des Fragebogens abbilden und die Skalenspannweite verbreitern. Somit ist die Skalenspannweite bei den Summen-Scores breiter als bei den Subskalen. Dies ist normalerweise eine gute Eigenschaft, weil man Floor- und Ceiling-Probleme vermeidet, aber das 15 %-Prinzip führt dann zu einem sehr hohen Responsekriterium.

Das folgende Szenario aus dem zitierten Schreiben verdeutlicht die Problematik des 15 %-Schwellenwerts beim PCS und MCS weiter und bezieht sich auf die Standard-Version des SF-36v2®:

Angenommen ein Patient hat seine Werte auf den physikalischen Subskalen körperliche Funktionsfähigkeit (physical functioning (PF)), körperliche Rollenfunktion (physical role functioning (RP)), körperliche Schmerzen (bodily pain (BP)) und allgemeine Gesundheit (general health (GH)) um 6,1; 6,0; 6,5 bzw. 7,3 Punkte verbessert. Dies sind etwas mehr als 15 % des Bereichs für jede Skala. Wenn die Werte auf den vier anderen Skalen (d.h. den mentalen Skalen) unverändert wären, würde die Verbesserung auf dem PCS 8,6 Punkte betragen. Dies ist eine numerisch größere Verbesserung beim PCS als auf jeder Subskala. Dennoch würde die Verbesserung auf der PCS-Skala nicht das 15 %-Prinzip erfüllen, obwohl die Verbesserung auf jeder der vier physikalischen Subskalen das 15 %-Prinzip erfüllt. Dies würde keine konsistente Interpretation der Ergebnisse ermöglichen.

Aus diesen Gründen ist Bjorner 2021 der Meinung, dass es unangemessen wäre, auf die Summen-Scores PCS und MCS das Prinzip des 15 %-Schwellenwerts anzuwenden. Die von den Entwicklern des SF-36v2® definierten Response-Schwellenwerte für PCS und MCS sind im Manual des SF-36v2® festgehalten (19) und betragen 3,4 Skalenpunkte auf der normbasierten Skala für PCS und 4,6 für MCS.

Novo Nordisk weist darauf hin, dass die von den Entwicklern definierten Response-Schwellenwerte für PCS und MCS erheblich von den Response-Schwellenwerten unter Anwendung des vom IQWiG festgesetzten 15 %-Schwellenwerts abweichen (diese sind 9.4 Skalenpunkte auf der normbasierten Skala für PCS und 9.6 für MCS). Der IQWiG-Response-Schwellenwert liegt beim PCS somit 176 % über dem von den Entwicklern definierten Schwellenwert (beim MCS sind es 109 %).

Aufgrund dieser erheblichen Erhöhung des Schwellenwerts bei PCS und MCS sieht Novo Nordisk eine wesentliche Gefahr, dass patientenrelevante Verbesserungen bei den beiden SF-36v2® Summenskalen PCS und MCS in der Nutzenbewertung in Zukunft nicht mehr erkannt werden können.

Janssen-Cilag GmbH

Beim SF-36v2 beispielsweise sind die Summenskalen für die Psychische Gesundheit (MCS, Mental Component Summary) und die Physische Gesundheit (PCS, Physical Component Summary) aus allen acht Subskalen konstruiert. Daher ist der Wertebereich breiter als bei den Subskalen (20).

18 Bjorner. Letter from QualityMetric about SF-36v2® Responder Criteria. 2021.

19 Maruish MEE. User's manual for the SF-36v2 Health Survey (3rd ed.). Lincoln, RI: QualityMetric Incorporated. 2011.

20 Bjorner J. QualityMetric's comment on the principle of using a blanket 15 % of the scale range as a response criterion for all patient-reported outcomes (PROs), in particular the implications for the interpretation of the physical component summary (PCS) and mental component summary (MCS) of the SF-36v2. 2021.

Als Konsequenz könnte die Anwendung eines 15 %-Schwellenwertes bedeuten, dass bei einer Verbesserung bzw. Verschlechterung in den Subskalen um einen Schwellenwert von 15 % in der Summenskala PCS zwar ebenfalls eine Verbesserung bzw. Verschlechterung demonstriert werden kann, die ihrerseits aber nicht den das Kriterium eines 15 %-Schwellenwertes erfüllt, obwohl dieses bei den Subskalen der Fall ist. Eine inkonsistente Interpretation der Ergebnisse wäre somit möglich (20).

Hinzu kommt, dass die theoretische Spannweite für die Summenskalen MCS und PCS in der Praxis im Gegensatz zu den Subskalen nicht beobachtet wird[20]. Für die Summenskalen sind zur Bestimmung der Skalenspannweite somit nicht die theoretischen Minima und Maxima, sondern die in der Praxis beobachteten Minima und Maxima heranzuziehen (20).

Eine Besonderheit tritt für Erhebungsinstrumente auf, die auf Populationen normiert sind. Je nach verwendeter Population bzw. dem Jahr der Normierung ergeben sich unterschiedliche Skalenspannweiten. Dies ist bei dem SF-36v2 der Fall. Für Responderanalysen, bei denen die zugrunde liegenden Daten des SF-36v2 auf die US-Population von 1998 normiert sind, müssten andere 15 %-Schwellenwerte herangezogen werden als für Responderanalysen, bei denen die zugrunde liegenden Daten des SF-36v2 auf die US-Population von 2009 normiert sind. Ein Vergleich von Ergebnissen ist somit unmöglich.

Bewertung

Die von den Stellungnehmenden vorgetragenen Responsekriterien basieren in der Regel auf Studien, die nicht dem aktuellen Stand der wissenschaftlichen Erkenntnis zur Bestimmung von MIDs entsprechen; es wird auf die Bewertung der Einwände zum Saint-George's Respiratory Questionnaire (SGRQ) auf den Seiten 11 und 12 verwiesen.

Die Einwände in Bezug auf die Normstichprobe und dem Phänomen, dass auf einzelnen Skalen eine Response beobachtet wird, nicht jedoch auf der Gesamtskala, stehen der Anpassung der Anlage II.6 zum 5. Kapitel der VerfO nicht entgegen.

Für den SF-36 entspricht die Auswertung mit einer Responseschwelle von ca. 10 Punkten der Umsetzung einer Responseschwelle von 15 % (Normstichprobe von 2009). Es wird auf die „Antworten auf häufig gestellte Fragen zum Verfahren der Nutzenbewertung", Unterkategorie „Fragen pharmazeutischer Unternehmen", Unterunterkategorie „Dossiererstellung" auf den Internetseiten des G-BA verwiesen.

Einwand

<u>European Organisation for Research and Treatment of Cancer – Quality of Life Questionnaire Cancer-30 (EORTC QLQ-C30) und Zusatzmodule</u>

<u>Merck Serono GmbH</u>

Eine von der AG Biostatistik des vfa vorgenommene Auswertung auf Basis von bisher in AMNOG-Verfahren akzeptierten MIDs zeigt zudem, dass die vorgeschlagene 15 %-Schwelle in ca. 90 % der Fälle zu einer Erhöhung der bisher etablierten und in der Anwendung akzeptierten MIDs führen würde. So würde z.B. für das in der Onkologie sehr häufig verwendete Erhebungsinstrument EORTC QLQ-C30 eine sehr deutliche Erhöhung der MID um 50 % resultieren (von bisher 10 auf 15). In diesem speziellen Fall und in allen anderen ähnlich gelagerten Fällen, wäre es nicht nachvollziehbar, warum etablierte MIDs, deren klinische Relevanz bereits belegt ist, im Rahmen der Nutzenbewertung nicht verwendet werden sollten.

Die EORTC beispielsweise hat es sich zum Ziel gemacht, MIDs für alle QLQ-C30 Skalen pro Indikationsgebiet zu etablieren (Musoro et al., 2019), erste Ergebnisse wurden bereits publiziert[21]. Die Ergebnisse solcher wissenschaftlich fundierten Arbeiten sollten

21 Musoro et al - MID for Interpreting EORTC QLQ-C30 Scores in Patients With Advanced Breast Cancer

Berücksichtigung und die daraus resultierenden Schwellenwerte Anerkennung in den Modulvorlagen finden.

Die Einführung einer generischen, universellen Schwelle von 15% Punkten der Skalenbreite würde zu einer regelhaften Erhöhung der Schwellenwerte führen, die zuvor als ausreichend galten, um den Zusatznutzen patienten- bzw. klinisch relevant zu evaluieren. Da höhere Messlatten schwerer zu überwinden sind, kann dies für patientenberichtete Auswertungen einen erschwerten methodischen Nachweis von Vor- und Nachteilen von Therapien bedeuten. Falls Verbesserungen oder Verschlechterungen des patientenzentrierten Befindens bestehen, könnten sie weniger häufig entdeckt werden. Dies stünde einer gerechten Bewertung im Rahmen der frühen Nutzenbewertung entgegen.

Des Weiteren ist der Schwellenwert von 15% Punkten infrage zu stellen, da die Skalierung des HRQoL Instruments ausschlaggebend sein kann für das Erreichen dieser generischen Schwelle. […]

Bayer Vital GmbH

[…] Jedenfalls ist der Ansatz einer exogen festgelegten 15% Schwelle der entsprechenden Skalenbreite nicht zielführend, da damit bereits validierte MIDs zu etablierten Instrumenten, die auch für entsprechende Studienplanungen herangezogen wurden, insofern diese niedriger ausfallen, hinfällig werden. Des Weiteren wird für eine Reihe von bereits erfolgten Nutzenbewertungen auf Basis dieser Instrumente mit den jeweilig angesetzten MIDs im Nachgang ihre Geltung in Frage gestellt. Beispielsweise würde für den in der Onkologie breit angewendeten EORTC QLQ-C30 eine nicht nachvollziehbare Erhöhung der MID um 50 % von bisher 10 auf 15 Punkte erfolgen, welche dessen bereits nachgewiesene klinische Relevanz unbegründet streitig machen würde.

Amgen GmbH

[…] Die Responderschwelle erscheint willkürlich und wissenschaftlich nicht ausreichend begründet. Gleichzeitig würde die vorgeschlagene Änderung zu einer Inkonsistenz bei solchen Verfahren in der frühen Nutzenbewertung führen, bei denen bisher instrumentspezifische MIDs < 15% als valide akzeptiert wurden. Hieraus ergeben sich deutliche Nachteile durch eine fehlende Vergleichbarkeit der Nutzenbewertungen verschiedener Wirkstoffe in einer Indikation, als auch durch die zusätzliche Hürde, über einen willkürlich höheren Schwellenwert einen signifikanten Vorteil zeigen zu müssen. Dies birgt das Risiko einer Benachteiligung gegenüber vorangegangenen Verfahren.

Beispielsweise wurde in der Indikation „Multiples Myelom“ bisher eine MID von ≥ 10 Punkten (entspricht 10 % der Skalenspannweite) zur Beurteilung einer patientenrelevanten Veränderung in den Skalen der Fragebögen EORTC QLQ-C30 und EORTC QLQ-MY20 als klinisch relevant akzeptiert (22, 23, 24, 25,

22 G-BA, Gemeinsamer Bundesausschuss 2018a. Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII – Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V – Carfilzomib (Neubewertung eines Orphan – Drugs nach Überschreitung der 50 Mio. Euro-Grenze); Datum: 15.02.2018

23 G-BA, Gemeinsamer Bundesausschuss 2018b. Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII – Beschlüsse über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V – Daratumumab (neues Anwendungsgebiet; Neubewertung eines Orphan Drugs nach Überschreitung der 50 Mio. Euro Grenze); Datum: 15.02.2018

24 G-BA, Gemeinsamer Bundesausschuss 2019. Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII – Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V Pomalidomid (neues Anwendungsgebiet: Kombinationstherapie Multiples Myelom); Datum: 05.12.2019

25 IQWiG, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2017. IQWiG-Berichte – Nr. 562: Daratumumab (multiples Myelom) – Nutzenbewertung gemäß § 35a SGB V; Datum: 13.11.2017

26, 27). Diese bereits akzeptierte MID basiert auf Validierungsstudien (28, 29, 30).

Auch der G-BA sieht eine MID von ≥ 10 Punkten beim EORTC QLQ-C30 und den Ergänzungsmodulen weiterhin als klinisch relevant an, wie in der Rubrik „Antworten auf häufig gestellte Fragen“ der G-BA Homepage zu finden ist:

„Eine Ausnahme bilden hierbei Auswertungen zum Fragebogen EORTC QLQ-C30 sowie den entsprechenden validierten krankheitsspezifischen Ergänzungsmodulen. Hierfür sind durch die pharmazeutischen Unternehmer in einer Übergangszeit (bis zum Inkrafttreten der angepassten Modulvorlagen) lediglich Auswertungen zur bisher akzeptierten MID von 10 Punkten im Dossier darzustellen.“ (31).

Ebenso bewertet das IQWiG in aktuellen Nutzenbewertungen (32, 33) die bisherige Responseschwelle von ≥ 10 Punkten für den EORTC QLQ-C30 und seine Ergänzungsmodule als angemessen und bezieht sich dabei auch auf die FAQs des G-BA:

„Zur Eignung der vom pU herangezogenen Responseschwelle von ≥ 10 Punkten für den EORTC QLQ-C30 und EORTC QLQ-CR29 (jeweilige Skalenspannweite 0-100) gilt Folgendes: Für den EORTC QLQ-C30 und seine Zusatzmodule wird die Auswertung mit der bisher akzeptierten Responseschwelle von 10 Punkten in bestimmten Konstellationen als hinreichende Annäherung an eine Auswertung mit einer 15 %-Schwelle (15 Punkte) betrachtet und für die Nutzenbewertung herangezogen. Unabhängig davon werden für eine Übergangszeit bis zum Inkrafttreten der angepassten Modulvorlagen für das Dossier primär Auswertungen mit der bisher akzeptierten Responseschwelle von 10 Punkten für den EORTC QLQ-C30 sowie alle Zusatzmodule des EORTC herangezogen (siehe FAQs des G-BA).“ (32).

Astellas Pharma GmbH

[…] Dessen ungeachtet führt die Etablierung der Responderschwelle von 15 % zu einer Inkonsistenz in der zukünftigen Bewertung des Zusatznutzens im Vergleich zu bereits abgeschlossenen Verfahren. Eine Auswertung von bisherigen AMNOG-Verfahren zeigt, dass es in ca. 90 % der Fälle zu einer Erhöhung der bisher validierten und seitens G-BA akzeptierten MIDs führen würde. Fraglich in diesem Zusammenhang ist auch, dass für ausgewählte

26 IQWiG, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2018. IQWiG-Berichte – Nr. 588: Carfilzomib (multiples Myelom) – Addendum zum Auftrag A17-38; Datum: 01.02.2018

27 IQWiG, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2019. IQWiG-Berichte – Nr. 814: Pomalidomid (multiples Myelom) – Nutzenbewertung gemäß § 35a SGB V; Datum: 12.09.2019

28 Cocks, K., King, M. T., Velikova, G. et al. 2011. Evidence-Based Guidelines for Determination of Sample Size and Interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. Journal of Clinical Oncology 29(1): 89-96.

29 Kvam, A. K., Fayers, P. M. and Wisloff, F. 2011. Responsiveness and minimal important score differences in quality-of-life questionnaires: a comparison of the EORTC QLQ-C30 cancer-specific questionnaire to the generic utility questionnaires EQ-5D and 15D in patients with multiple myeloma. European Journal of Haematology, 87(4): 330-337.

30 Sully, K., Trigg, A., Bonner, N. et al. 2019. Estimation of minimally important differences and responder definitions for EORTC QLQ-MY20 scores in multiple myeloma patients. European Journal of Haematology, 103(5): 500-509.

31 G-BA, Gemeinsamer Bundesausschuss 2021. Antworten auf häufig gestellte Fragen zum Verfahren der Nutzenbewertung – „Wie soll, vor dem Hintergrund der Veröffentlichung des Methodenpapiers 6.0 des IQWiG am 5. November 2020, derzeit in der Dossiererstellung mit der Bestimmung von klinischen Relevanzschwellen bei komplexen Skalen umgegangen werden?“

32 IQWiG, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2021a. IQWiG-Berichte – Nr. 1144: Pembrolizumab (Kolorektalkarzinom mit MSI-H oder dMMR) – Nutzenbewertung gemäß § 35a SGB V; Datum: 29.06.2021 [Zugriff: 12.07.2021]. URL: https://www.g-ba.de/downloads/92-975-4607/2021-04-01_Nutzenbewertung-IQWiG_Pembrolizumab_D-653.pdf.

33 IQWiG, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2021b. IQWiG-Berichte – Nr. 1061: Atezolizumab (hepatozelluläres Karzinom) – Nutzenbewertung gemäß § 35a SGB V; Datum: 25.02.2021 [Zugriff: 12.07.2021]. URL: https://www.iqwig.de/download/a20-97_atezolizumab_nutzenbewertung-35a-sgb-v_v1-0.pdf.

Endpunkte, wie dem EORTC QLQ-C30, bereits von der Responderschwelle von 15 % abgewichen wurde. […]

Bewertung

Es ist richtig, dass die Definition der Responseschwelle von 15 % der Skalenspannweite dazu führt, dass zum Teil bisher verwendete Responsekriterien nicht mehr berücksichtigt werden.

Die Ausführungen der Stellungnehmenden dazu, welche Auswirkungen die 15 %-Schwelle insbesondere für die in der Onkologie häufig angewendeten EORTC-Fragebögen (z.B. EORTC QLQ-C30) hat, sind jedoch falsch. Denn die Stellungnehmenden argumentieren mit theoretischen Überlegungen einer Erhöhung des Responsekriteriums um 50 % (von 10 auf 15 Punkte), ohne den Aufbau des Fragebogens EORTC QLQ-C30 und die damit verbundenen praktischen Auswirkungen eines geänderten Responsekriteriums zu berücksichtigen, was nachfolgend näher erläutert wird:

- Der Fragebogen EORTC QLQ-C30 stellt eine Sammlung von Skalen zu Symptomen und Funktionen dar. Er enthält insgesamt 15 Skalen: 9 Skalen zu verschiedenen Symptomen (z.B. Luftnot), 5 Funktionsskalen und eine Skala zum globalen Gesundheitszustand.
- Die einzelnen Skalen können ein oder mehrere Items, d.h. konkrete Fragen, enthalten (bis zu 4). Für jedes Item (Frage) sind verschiedene fest vorgegebene Antworten möglich, z.B. keine Symptome oder starke Symptome. Das Ergebnis für eine Skala (z.B. für das Symptom Luftnot) entspricht der Summe der Ergebnisse aller zur Skala gehörenden Items (Fragen).
- Der Punktebereich wird für jede Skala (z.B. das Symptom Luftnot) jeweils auf einen Bereich von 0 bis 100 Punkte normiert, unabhängig davon, wie viele Items die jeweilige Skala enthält. Das bedeutet: die geringste mögliche Punktzahl der Skala wird einem Wert von 0 zugeordnet, die höchst mögliche einem Wert von 100. Das Responsekriterium (10 Punkte oder 15 %) bezieht sich auf diesen normierten Bereich von 0 bis 100 Punkte. 15 % entsprechen also für jede Skala des Fragebogens EORTC QLQ-C30 einer Veränderung um 15 Punkte.
- Ein Großteil der 15 Skalen besteht nur aus 1 Item mit 4 Antwortmöglichkeiten. Das bedeutet für die Normierung: Die erste Antwortmöglichkeit, z.B. „keine Symptome“, wird dem Wert 0 zugeordnet, die letzte Antwortmöglichkeit, z.B. „starke Symptome“, dem Wert 100. Die beiden dazwischenliegenden Antwortmöglichkeiten (z.B. „wenige Symptome“ bzw. „moderate Symptome“) werden entsprechend den Werten 33,33 bzw. 66,67 auf der normierten Skala zugeordnet (gleicher Abstand zwischen den 4 verschiedenen Antwortmöglichkeiten). Für diese Skalen bedeutet also jegliche Veränderung der Antwort (z.B. von „keine Symptome“ bei der ersten Messung zu „wenige Symptome“ bei der Folgemessung) eine Veränderung des Punktwerts um mindestens 33,33 Punkte. Veränderungen unterhalb eines Punktwerts von 33,33 Punkten sind nicht möglich. Ob also als Responsekriterium 10 Punkte (bisher verwendetes Responsekriterium) oder 15 % = 15 Punkte (neues Responsekriterium) angewendet wird, ist für diese Skalen ohne jegliche praktische Auswirkung.
- Für den Großteil der Skalen des EORTC QLQ-C30-Fragebogens, nämlich 13 der 15 Skalen, hat die Änderung des Responsekriteriums von 10 auf 15 Punkte (entsprechend 15 % der Skalenspannweite) keinerlei praktische Konsequenz. Für die anderen beiden Skalen ergibt sich jeweils nur eine höchstens geringe Abweichung zwischen den beiden Responsekriterien (maximal Unterschied von einem Antwortpunkt bei einem einzelnen Item).
- Das zuvor Gesagte trifft in ähnlicher Größenordnung auch für die krankheitsspezifischen EORTC-Fragebögen zu, wie z.B. den Fragebogen EORTC QLQ-

LC13. Hier sind die Responsekriterien 10 Punkte und 15 % bei 9 der insgesamt 10 Skalen dieses Fragebogens inhaltlich deckungsgleich und ein Wechsel des Responsekriteriums von 10 Punkte auf 15 % daher ohne praktische Auswirkungen.

Zusammenfassend ergibt sich entgegen der Einschätzung der Stellungnehmenden aufgrund des Aufbaus der EORTC-Fragebögen in den weitaus überwiegenden Fällen kein Unterschied bei der Verwendung eines Responsekriteriums von 10 Punkten oder 15 % (entsprechend 15 Punkten). Es ist aus Praktikabilitätsgründen sinnvoll, auch zukünftig einheitlich für alle Skalen dieser Fragebögen das bislang verwendete Responsekriterium 10 Punkte zu akzeptieren, da zumeist kein inhaltlicher Unterschied zwischen den beiden Responsekriterien besteht. Auch zukünftig sind daher für EORTC-Fragebögen nur Auswertungen zum Responsekriterium 10 Punkte im Dossier darzustellen.

Es wird auf die „Antworten auf häufig gestellte Fragen zum Verfahren der Nutzenbewertung“, Unterkategorie „Fragen pharmazeutischer Unternehmen“, Unterunterkategorie „Dossiererstellung“ auf den Internetseiten des G-BA verwiesen.

Einwand

Hamilton Rating Scale for Depression (HAMD), Montgomery–Åsberg Depression Rating Scale (MADRS), Positive and Negative Syndrome Scale (PANSS)

Deutsche Gesellschaft für Psychiatrie und Psychotherapie, Psychosomatik und Nervenheilkunde e. V.

Die in der Psychiatrie gebräuchlichen Fremdratingskalen nutzen in der Regel nur 50 %, im Einzelfall 60 % der Spannbreite nach oben aus. Ein 15 %-Schwellenwert (von der Spannweite) für eine klinische Verbesserung stellt daher aus Sicht der DGPPN eine zu hohe klinische Signifikanzschwelle dar. […]

Weder aus den zur Verfügung gestellten Dokumenten noch dem IQWiG-Methodenpapier 6.0 wird deutlich, ob mit „patientenberichteten Endpunkten“ nur Selbstratings (und damit „patientenberichtete Endpunkte“ im engeren Sinne) gemeint sind oder ob auch Fremdratings – die üblicherweise in Arzneimittelprüfungen in der Psychiatrie nach wie vor den Goldstandard darstellen und auf die sich jede Zulassung durch EMA oder FDA stützt – umfasst sind. Ob Fremdratings von G-BA und IQWiG überhaupt oder ausreichend diskutiert wurden, ist ebenfalls unklar.

Wenn auch Fremdratings in die Schwellenwertbetrachtung mit einbezogen sind, so ist die o. g. Änderung eindeutig kritisch zu betrachten wie am folgenden Beispiel deutlich wird: Die HAMD als (neben der MADRS) von EMA und FDA als Standardskala betrachteter primärer Endpunkt hat eine Spannweite von 52 Punkten. 15 % wären 7,8 Punkte. Dies als Minimum für eine klinisch relevante Veränderung zu betrachten, ist unrealistisch. Eine Veränderung von 7 Punkten ist klinisch hochrelevant. Ein Absinken von 23 auf 16 Punkte stellt eine Veränderung von einer schweren auf eine leichte Depression dar. Ein Abfall von 14 auf 7 Punkte stellt eine Remission einer leichten Depression dar. Bei der MADRS verhält es sich ähnlich. Das Problem mit diesen Skalen ist, dass die Spannweite in der klinischen Praxis nie ausgeschöpft wird. Schon ein Patient mit 30 Punkten auf der HAMD muss als schwerst depressiver Patient gelten. Ein Patient mit 40 Punkten wird sehr wahrscheinlich nicht in eine Studie eingeschlossen.

Für die PANSS als Standardinstrument in der Evaluation von Antipsychotika gilt Vergleichbares. Die Skala hat eine Spannweite von 210 Punkten. 15 % Besserung entsprächen 31,5 Punkte. Das ist als klinisches Signifikanzkriterium ebenfalls zu hoch, wenn auch eher zu erreichen als auf einer der gängigen Depressionsskalen. Nach den Arbeiten von Stefan Leucht entsprechen 20 % Besserung auf der PANSS „minimally improved“ auf der CGI. Bei einer Baseline von 80 Punkten auf der PANSS wären 20 % 16 Punkte, was bereits klinisch signifikant ist. In vielen Schizophreniestudien gelten 30 % Besserung als Responsekriterium (bei Ersterkrankten auch 50 %), aber „Response“ ist mehr als eine klinisch relevante Veränderung. Es ist

vom Ausgangsniveau abhängig: Ein Patient mit einem Score von 65 Punkten ist nicht mehr besonders schwer krank, aber er dürfte immer noch signifikant beeinträchtigt sein. Ein Abfall des PANSS-Scores um 31,5 Punkte macht ihn aber zu einem Gesunden. Ein Patient mit einem Ausgangsscore von 100 Punkten wäre schwer krank, eine Besserung um 30 Punkte würde eine „Response“, also eine erhebliche Besserung darstellen, das ist sicher mehr als eine „klinische relevante“ Veränderung.

Zusammengenommen erachtet die DGPPN daher einen 15 %-Schwellenwert von der Skalenspannweite als eine zu hohe klinische Signifikanzschwelle für eine klinische Verbesserung. Vorgeschlagen wird daher, dass sich der 15 %-Schwellenwert auf die Spanne der Baseline der konkreten Population in der Studie beziehen soll oder auf die Spannweite, die die untersuchte Studienpopulation allgemein (basierend auf anderen Studien oder Metaanalysen) charakterisiert.

Bewertung

Grundsätzlich können von der Anpassung der Anlage II.6 zum 5. Kapitel der VerfO auch Fremdratings umfasst sein. Die Anpassung gilt generell für patientenrelevante Endpunkte, die mit (komplexen) Skalen erhoben werden, und für die kleine Änderungen beschrieben werden sollen. In diesen Fällen ist es in besonderer Weise notwendig, neben der statistischen Signifikanz der Effekte die Relevanz der beobachteten Wirkungen der untersuchten Interventionen zu bewerten. Dabei wird die Interpretation der Daten durch die Komplexität der Skalen erschwert.

Im Gegensatz dazu ist die Interpretation von Ergebnissen, die mit einfachen Skalen erhoben werden (z. B. zu Häufigkeit eines Symptoms), in der Regel direkt möglich (z. B. die Einschätzung der Relevanz einer Änderung von 0,2 Symptomepisoden pro Woche).

Patientenrelevante Endpunkte, die zwar mit Skalen erhoben werden, für die aber eher große Änderungen betrachtet werden sollen, sind von der Anpassung der Anlage II.6 nicht betroffen. Das gilt für die von den Stellungnehmenden beschriebenen Anwendungen der in der Psychiatrie gebräuchlichen Skalen wie HAMD, MADRS oder PANSS. Die für diese Skalen gebräuchlichen Schwellen für Response und Remission stellen ausgeprägte Änderungen der Krankheitslast dar. Diese Kriterien wurden auch bisher nicht im Sinne einer MID interpretiert. Weitere Beispiele für Skalen, deren Standardauswertung große Änderungen der Symptomlast und eben keine MID abbilden, sind der PASI oder der EASI (z.B. 90 % Änderung oder 100 % Änderung).

<u>Weitere Anmerkungen</u>

Einwand

> <u>Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF); Deutsche Gesellschaft für Pharmazeutische Medizin e. V.</u>

[…] Man kann davon ausgehen, dass die im Kontext von klinischen Studien genutzten Instrumente zur Erfassung der PRO in den meisten, wenn auch nicht allen Fällen, den publizierten Vorgaben von EMA und FDA entsprechen.

Vor allem in entsprechenden Publikationen des britischen NICE sind jedoch auch Defizite der jeweils genutzten Instrumente in einigen Indikationen veröffentlicht, sei es, dass die Validierung zum Zeitpunkt des Einsatzes noch unvollständig war oder weitere Daten fehlten. Auch aus den nationalen Diskussionen im Zusammenhang mit Verfahren zur frühen Nutzenbewertung sind die immer wieder auftretenden Differenzen um die Validität genutzter Instrumente wie auch um die Festlegung einer Responseschwelle bekannt.

Dies bedeutet aber nicht, dass die wissenschaftliche Grundlage aller in klinischen Studien genutzten Instrumente zur Erfassung der PRO in Frage gestellt werden muss. Auch das IQWIG hat in seinen Bewertungen sowohl die Wahl von Instrumenten als geeignet erachtet als auch

die gesetzten Responseschwellen akzeptiert. Natürlich ist nachvollziehbar, zur allgemeinen Vereinfachung der Bewertungen und zur Vermeidung einer ergebnisgesteuerten Berichterstattung, für alle Instrumente eine einheitliche Responseschwelle festzulegen. Diesen Ansatz hat das IQWiG in der letzten Version seines Methodenpapiers (Version 6.0) bereits vorgestellt.

Laut IQWiG soll eine feste Responseschwelle von 15 % der Skalenspannweite nun „Klarheit für die Hersteller schaffen und willkürliche Responderanalysen auf Basis nicht nachvollziehbarer Responderdefinitionen verhindern".

Hintergrund für diese Entscheidung ist die Einschätzung des IQWIG, dass viele wissenschaftliche Untersuchungen zur Ermittlung einer MID nicht mehr den heutigen methodischen Ansprüchen genügen oder die angewendete Methodik in den wissenschaftlichen Publikationen nicht hinreichend beschrieben wird. In dieser Situation hat das IQWiG einen neuen Ansatz entwickelt, der es ermöglichen soll, auf einfachem Wege Schwellen zu ermitteln, die hinreichend sicher einen bedeutsamen Bereich abgrenzen.

Allerdings interpretiert die DGPharMed dies nicht so, dass bei allen Bewertungen nur noch die vom IQWIG gesetzte Schwelle von 15% gilt. Unter dem neuen Kriterium einer Responseschwelle von 15 % des Skalenrangs werden eine Reihe von bisher akzeptierten, wissenschaftlich fundierten „MIDs" nicht mehr berücksichtigt.

Sofern im Rahmen eines Modul 4 eine wissenschaftlich fundierte Herleitung einer anderen, a priori bereits im SAP einer klinischen Studie festgelegten Responseschwelle vorgelegt wird, sollte dies ebenfalls oder stattdessen für die Bewertung akzeptiert werden. Wann eine wissenschaftliche Herleitung den heutigen methodischen Ansprüchen genügt oder die angewendete Methodik in den wissenschaftlichen Publikationen hinreichend beschrieben wird, sollte allerdings Thema der wissenschaftlichen Fachkreise und nicht nur einer einzelnen nationalen Institution sein.

IQVIA Commercial GmbH & Co. OHG

IQWiG attestiert seinem Schwellenwert „Praxistauglichkeit“

Die Zwischenbilanz des IQWiG nach den ersten Anwendungen des 15% Responsekriteriums fällt positiv aus – was zumindest in der Wortwahl der Bilanz verwundert, da im Sinne des Verfahrens andere Gütekriterien als eine reine Praktikabilität und Handhabbarkeit als Bewertungsmaßstab dienen sollten. Das Augenmerk sollte insbesondere auf der Sensitivität für kleine aber für den Patienten bedeutsamen Veränderungen liegen.

Positiv aus unserer Sicht war allein die transparente methodische Auseinandersetzung mit den Spezifika einiger Skalen:

- So ergeben sich bei nahezu allen EORTC Skalen keine Unterschiede bei der Anwendung des 15%-Responsekriteriums gegenüber der etablierten MID von 10 Punkten – ein Umstand, der sich aus dem Scoring-Algorithmus der EORTC-Skalen ergibt (34) und auch ohne die auf klinischen Daten basierende Gegenüberstellung einleuchtet (35). Nicht nachvollziehbar allerdings sind die Inkonsistenzen bei der Akzeptanz des 10-

34 European Organisation for Research and Treatment of Cancer (EORTC) 2001. EORTC QLQ-C30 Scoring Manual.

35 Böhm D. 2020. MID – One Size Fits All? GMDS-ATF Workshop | 5. November 2020.

Punktekriteriums über unterschiedlich onkologische Nutzenbewertungen hinweg (36, 37).

- Probleme bei der Definition der Skalenspannweite wie beispielsweise beim SF-36 waren vorab nicht antizipiert worden, wurden jedoch vom IQWiG methodische aufgearbeitet (38, 39); inwieweit auch für andere Skalen die Spannweite einer „Scale in practice“ anstelle einer „Theoretical scale“ Basis für die Berechnung der Skalenspannweite verwendet wird bleibt fraglich.
- positiv ebenfalls die differenzierte Betrachtung durch den G-BA aufgrund der Gleichbewertung einzelner Verfahren im Anwendungsgebiet {z.B. (40, 38)}.

Uns stellt sich allerdings die Frage, als wie „praxistauglich“ sich das 15%-Kriterium angesichts einer angestrebten Harmonisierung der Nutzenbewertung auf europäischer Ebene erweisen wird. Die Anforderung an die Auswertung und Ergebnisdarstellung von Skalen sind als deutscher Alleingang zu sehen und finden weder in den regulatorischen Guidelines noch in den Guidelines der EUNetHTA ihre Entsprechung (41, 42, 43, 44). Welche Empfehlungen wird der G-BA in der frühen Beratung aussprechen?

Ecker + Ecker GmbH

2 Die Bedeutung von patientenberichteten Endpunkten (PRO) wird durch dieses Vorgehen in Frage gestellt

Für die Interpretation von patientenberichteten Endpunkten stellen MCID einen zentralen Bestandteil dar. Sie dienen dazu, eine für den Patienten spürbare und wichtige Veränderung sensitiv von trivialen Änderungen zu trennen und sollen die Patientenperspektive in die

36 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) 2021a. Atezolizumab (hepatozelluläres Karzinom) – Nutzenbewertung gemäß § 35a SGB V. Verfügbar unter: https://www.g-ba.de/downloads/92-975-4212/2020-12-01_Nutzenbewertung-IQWiG_Atezolizumab_D-603.pdf, abgerufen am: 14.07.2021.

37 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) 2021b. Carfilzomib (multiples Myelom) – Nutzenbewertung gemäß § 35a SGB V. Verfügbar unter: https://www.g-ba.de/downloads/92-975-4382/2021-01-15_Nutzenbewertung-IQWiG_Carfilzomib_D-617.pdf, abgerufen am: 14.07.2021.

38 Gemeinsamer Bundesausschuss (G-BA) 2021. Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII – Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V Secukinumab (Neubewertung aufgrund neuer wissenschaftlicher Erkenntnisse (PsoriasisArthritis)). Verfügbar unter: https://www.g-ba.de/downloads/40-268-7322/2021-02-18_AM-RL-XII_Secukinumab_D-576_TrG.pdf, abgerufen am: 14.07.2021.

39 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) 2021c. Filgotinib (rheumatoide Arthritis) – Nutzenbewertung gemäß § 35a SGB V. Verfügbar unter: https://www.g-ba.de/downloads/92-975-4088/2020-10-15_Nutzenbewertung-IQWiG_Filgotinib_D-590.pdf, abgerufen am: 14.07.2021.

40 Gemeinsamer Bundesausschuss (G-BA) 2021b. Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM RL): Anlage XII – Anlage XII – Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V Ipilimumab (Neues Anwendungsgebiet: Nicht-kleinzelliges Lungenkarzinom, Kombination mit Nivolumab und platinbasierter Chemotherapie, Erstlinie). Verfügbar unter: https://www.g-ba.de/downloads/40-268-7571/2021-06-03_AM-RL-XII_Ipilimumab_D-629_TrG.pdf, abgerufen am: 22.07.2021.

41 European Medicines Agency (EMA) 2016. Appendix 2 to the guideline on the evaluation of anticancer medicinal products in man. Verfügbar unter: https://www.ema.europa.eu/en/documents/other/appendix-2-guideline-evaluation-anticancer-medicinal-products-man_en.pdf, abgerufen am: 22.07.2021.

42 European Network for Health Technology Assessment (EUNetHTA) 2013. Guideline - Endpoints for relative effectiveness assessment of pharmaceuticals: Health-related Quality of Life and Utility Measures. Verfügbar unter: https://www.eunethta.eu/wp-content/uploads/2013/01/Health-related-quality-of-life.pdf, abgerufen am: 22.07.2021.

43 European Network for Health Technology Assessment (EUNetHTA) 2015. Guideline - Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints. Verfügbar unter: https://www.eunethta.eu/wp-content/uploads/2018/02/WP7-SG3-GL-clin_endpoints_amend2015.pdf, abgerufen am: 22.07.2021.

44 Food and Drug Administration (FDA) 2009. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Verfügbar unter: https://www.fda.gov/media/77832/download, abgerufen am: 22.07.2021.

Interpretation des Instrumentes einbinden. Aus diesem Grund werden sie nicht willkürlich gewählt, sondern idealerweise anhand von klinischen Kriterien abgeleitet, welche die Patientenperspektive berücksichtigen sollen und durchaus zu je nach Indikation unterschiedlichen MCID führen können. Beispiele für solche klinischen Kriterien sind patientenberichtete Anker. Das skalen- und indikationsübergreifende 15 %-Kriterium beruht auf keiner derartigen klinischen Validierung und negiert daher die Patientenperspektive für die Interpretation von patientenberichteten Endpunkten.

So ignoriert das 15%-Kriterium Veränderungen der Lebensqualität oder anderer patientenrelevanter Parameter, die unterhalb der 15%-Schwelle liegen, aber für die Patienten spürbar und wichtig sind. Umgekehrt kann das Kriterium zur Anwendung einer zu niedrigen Relevanzschwelle führen, die fälschlicherweise einen Patientennutzen suggeriert. So würde die Anwendung auf die Symptom Severity Scale des Uterine Fibroid Symptom and Quality of Life Questionnaire (UFS-QoL) zu einer Relevanzschwelle von 15 Punkten führen, obwohl das IQWiG davon ausgeht, dass selbst 17 Punkte keine für die Patienten spürbare Änderung bedeuten [45]. Zudem werden Analysen, welche die Stabilität eines erreichten Gesundheitszustandes nachweisen wollen (beispielsweise die Zeit bis zur Verschlechterung der Lebensqualität) durch die Anwendung des 15%-Kriteriums erschwert, da kleinere Veränderung, die für den Patienten eine relevante Verschlechterung darstellen könnten, ignoriert werden.

Aufgrund der hohen Relevanz, welche der Lebensqualität in der frühen Nutzenbewertung zukommen soll, ist es nicht nachvollziehbar, warum für die Patienten spürbare Effekte ignoriert werden, nur weil sie nicht die Grenze von 15% der Skalenspannweite erreichen, und Änderungen oberhalb von 15% der Skalenspannweite pauschal als relevant eingestuft werden sollen. Durch die Abkopplung des Relevanzkriteriums von für die Patienten spürbaren Kriterien besteht die Gefahr, einen patientenrelevanten Nutzen nicht zu erkennen oder einen für die Patienten wichtigen Schaden nicht als solchen zu detektieren. Der Wert patientenberichteter Endpunkte wird durch ein solches Vorgehen in Frage gestellt.

3 Die Praktikabilität des 15%-Kriteriums bleibt fraglich

Das 15%-Kriterium soll als universell anwendbares Kriterium in die Dossiervorlage aufgenommen werden. Doch schon jetzt zeichnet sich ab, dass das Kriterium nicht auf alle Skalen anwendbar ist. So lässt sich das 15%-Kriterium nicht sinnvoll auf die theoretisch mögliche Skalenspannweite der Summenskalen des SF-36 anwenden, da in der Praxis die möglichen Maximalwerte nicht erreicht werden. Auch das IQWiG schlägt hier die Anwendung des Kriteriums auf die in einer Normstichprobe ermittelte Skalenspannweite vor, also 10 Punkte [46]. Das Problem kann aber auch für andere Instrumente bestehen und es bleibt offen, anhand welcher Kriterien entschieden werden soll, auf welche Skalenspannweite das 15%-Kriterium angewendet werden soll.

Bewertung

Die Studien zur Bestimmung von MIDs entsprechen in der Regel nicht dem aktuellen Stand der wissenschaftlichen Erkenntnis; es wird auf die Bewertung der Einwände zum Saint-George's Respiratory Questionnaire (SGRQ) auf den Seiten 11 und 12 verwiesen.

Mit Anpassung der Anlage II.6 zum 5. Kapitel der VerfO soll die Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen entsprechend des Methodenpapiers 6.0 des IQWiG sowie den Vorgaben der Modulvorlage erfolgen. Somit

45 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Magnetresonanztomografiegesteuerte hochfokussierte Ultraschalltherapie zur Behandlung des Uterusmyoms (Addendum zu den Aufträgen E14-04 und E14-05). 2017.

46 Schürmann C. Methodische Aspekte bei der Analyse von Daten zur Lebensqualität in Nutzenbewertungen. IQWiG im Dialog 2021; virtuelle Konferenz 2021.

entspricht beispielsweise eine präspezifizierte Responderanalysen unter Verwendung eines Responsekriteriums < 15 % der Skalenspannweite des verwendeten Erhebungsinstruments nicht dem Methodenpapier 6.0 des IQWiG sowie den Vorgaben der Modulvorlage.

Für die EORTC-Fragebögen wird auf die Ausführungen auf den Seiten 17 und 18 sowie auf die „Antworten auf häufig gestellte Fragen zum Verfahren der Nutzenbewertung“, Unterkategorie „Fragen pharmazeutischer Unternehmen“, Unterunterkategorie „Dossiererstellung“ auf den Internetseiten des G-BA verwiesen.

Für den SF-36 entspricht die Auswertung mit einer Responseschwelle von ca. 10 Punkten der Umsetzung einer Responseschwelle von 15 % (Normstichprobe von 2009).

2.3.2 Methodische Vorgehen zur Ableitung der Responseschwelle von 15 %, aktuelle wissenschaftliche Diskussionen zur Bewertung von MIDs und etablierte Standards

Einwand

Novo Nordisk Pharma GmbH

[…] Zur Herleitung der Responseschwelle von 15 % bezog sich das IQWiG im Entwurf der Allgemeinen Methoden Version 6.0 von 2019 auf „eine Sichtung von systematischen Übersichtsarbeiten“. Hier blieb unklar, ob diese Sichtung vollständig war, da sich die beiden einzigen vom IQWiG exemplarisch angegebenen Quellen (47, 48) lediglich auf ausgewählte und sehr spezifische Symptome bezogen.

In der finalen Version der Allgemeinen Methoden Version 6.0 von 2020 (49) wurde die Auswahl der systematischen Übersichtsarbeiten vom IQWiG von zwei auf insgesamt acht Quellen erweitert (50, 51, 52, 53, 54, 55).

Außer der Angabe einer systematisch recherchierten Sichtung von systematischen Übersichtsarbeiten gibt es auch in der finalisierten Version der Methoden 6.0 vom IQWiG keine näheren methodischen Angaben zu dieser Literaturrecherche, unklar bleiben auch die verwendeten Selektionskriterien. […]

In der Gesamtschau der vom IQWiG selektierten indikationsabhängigen systematischen Übersichtsarbeiten ergeben sich aus Sicht von Novo Nordisk keine Anhaltspunkte dafür, dass die Intention der Autoren darin gelegen hatte, indikationsübergreifende generell gültige MIDs/MICDs für PRO-Skalen zu entwickeln. Den Autoren ging es primär darum den Stand der MID/MICD-Entwicklung in den spezifischen Anwendungsgebieten zu dokumentieren, Evidenzlücken zu identifizieren oder Anleitung für die Anwendung der vorhandenen MIDs/MICDs zu geben.

Die einzige vom IQWiG selektierte Indikations-unabhängige Arbeit liefert Jayadevappa 2017 (54). Hier werden MIDs/MICDs für generische und krankheitsspezifische Skalen-Instrumente zur Messung von PRO´s untersucht. Das Ergebnis dieser Arbeit ist aber keine allgemeine Responseschwelle für MIDs/MICDs. Die Autoren kommen vielmehr zu dem Schluss, dass es für alle der untersuchten PRO-Skalen, kein einheitliches Maß für eine MID oder MCID gibt.

47 Nordin A, Taft C, Lundgren-Nilsson A, Dencker A. Minimal important differences for fatigue patient reported outcome measures-a systematic review. BMC Med Res Methodol. 2016;16:62.

48 St-Pierre C, Desmeules F, Dionne CE, Fremont P, MacDermid JC, Roy JS. Psychometric properties of self-reported questionnaires for the evaluation of symptoms and functional limitations in individuals with rotator cuff disorders: a systematic review. Disabil Rehabil. 2016;38(2):103-22.

49 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Allgemeine Methoden. Version 6.0 vom 05.11.2020.2020. Available from: https://www.iqwig.de/methoden/allgemeine-methoden_version-6-0.pdf?rev=180500.

50 Alma H, de Jong C, Tsiligianni I, Sanderman R, Kocks J, van der Molen T. Clinically relevant differences in COPD health status: systematic review and triangulation. Eur Respir J. 2018;52(3).

51 Doganay Erdogan B, Leung YY, Pohl C, Tennant A, Conaghan PG. Minimal Clinically Important Difference as Applied in Rheumatology: An OMERACT Rasch Working Group Systematic Review and Critique. J Rheumatol. 2016;43(1):194-202.

52 Ebrahim S, Vercammen K, Sivanand A, Guyatt GH, Carrasco-Labra A, Fernandes RM, et al. Minimally Important Differences in Patient or Proxy-Reported Outcome Studies Relevant to Children: A Systematic Review. Pediatrics. 2017;139(3).

53 Hao Q, Devji T, Zeraatkar D, Wang Y, Qasim A, Siemieniuk RAC, et al. Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation. BMJ Open. 2019;9(2):e028777.

54 Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life-a systematic review. J Clin Epidemiol. 2017;89:188-98.

55 Ousmen A, Touraine C, Deliu N, Cottone F, Bonnetain F, Efficace F, et al. Distribution- and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review. Health Qual Life Outcomes. 2018;16(1):228.

Nach Einschätzung von Jayadevappa 2017 hängt die MID-Schätzung von dem Kontext der Erkrankung, dem Schweregrad der Erkrankung, den Merkmalen der Studienpopulation, der interessierenden Untersuchungseinheit (Individuum oder Gruppe), den beobachteten Ausgangswerten und der Veränderung der Skalenwerte innerhalb der Beobachtungszeit ab.

Nichtsdestotrotz ergab sich für das IQWiG aus diesen acht beschriebenen Quellen seiner fokussierten - nicht systematischen - Literaturrecherche das regelhafte Vorgehen, bei der Beurteilung von Skalen ein einheitliches Responsekriterium von 15 % der theoretisch möglichen Skalenspannweite anzuwenden.

Die Ableitung des pauschalen Responsekriteriums von 15 % ist aus Sicht von Novo Nordisk wissenschaftlich und evidenzbasiert weiterhin nicht nachvollziehbar und ist daher abzulehnen ist. […]

Auch die vom IQWiG in den Allgemeinen Methoden Version 6.0 zitierte Arbeit von Revicki et al. 2008 (56) beschreibt, dass eine MID für ein PRO-Instrument kein unveränderbarer fester Wert ist, sondern populations- und kontextabhängig ist. Zur Findung von sinnvollen Responseschwellen (MIDs) sollten nach Revicki et al. 2008 ankerbasierte Methoden verwendet werden, die sowohl patientenspezifische, klinische und krankheitsspezifische Eigenschaften heranziehen.

Novo Nordisk schließt sich dieser Empfehlung an. Insbesondere sieht Novo Nordisk die Verwendung von patienten- und klinisch basierten Ankern als wesentliches Element an, um die Patientenrelevanz von Responseschwellen (MIDs) sicherzustellen.

MSD Sharp & Dohme GmbH

Trotz anhaltender Diskussion der internationalen Fachgemeinschaften zu verbesserten Standards für die Bewertung von MID, [57, 58, 59, 60, 61, 62] hat das IQWiG in seinen „Allgemeinen Methoden“ Version 6.0 (veröffentlicht am 05. November 2020) überraschend einen eigenständige Responderschwelle mit einen Schwellenwert von 15 % als plausibel definiert. Diese generische Schwelle von 15% basiert auf der angegebenen Spannweite der erhobenen MID in acht Literaturstellen und leitet sich methodisch nicht genügend fundiert ab. So wurden beispielhaft niedrige MID Werte als unrealistisch abgestuft und nicht in Betrachtung gezogen. Zudem verwiesen selbst die Autoren der zitierten systematischen Übersichtsarbeiten auf erhebliche Limitationen hin, welche die Übertragbarkeit der

56 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61(2):102-9.

57 Coon, C. D., & Cook, K. F. (2018). Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Quality of Life Research, 27(1), 33-40. https://link.springer.com/article/10.1007%2Fs11136-017-1616-3

58 Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of clinical epidemiology, 61(2), 102-109.. https://www.jclinepi.com/article/S0895-4356(07)00119-9/fulltext

59 U.S. Department of Health and Human Services FDA, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health. (2009). Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. https://www.fda.gov/media/77832/download

60 Musoro, Z. J., Hamel, J. F., Ediebah, D. E., Cocks, K., King, M. T., Groenvold, M., ... & Coens, C. (2018). Establishing anchor-based minimally important differences (MID) with the EORTC quality-of-life measures: a meta-analysis protocol. BMJ open, 8(1), e019117. https://bmjopen.bmj.com/content/10/2/e032112

61 Devji, T., Carrasco-Labra, A., Qasim, A., Phillips, M., Johnston, B. C., Devasenapathy, N., ... & Guyatt, G. H. (2020). Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. Bmj, 369. https://www.bmj.com/content/369/bmj.m1714

62 Peipert JD, Cella D. (2020). Rapid Response to “Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study”. Bmj, 369. https://www.bmj.com/content/369/bmj.m1714/rapid-responses

Ergebnisse selbst innerhalb des Therapiegebiets als unzureichend erscheinen lässt. [z.B. [63, 64]]

So sehr die Diskussion zu verbesserten Standards für die Bewertung von MID für patientenberichtete Endpunkte zu begrüßen ist, so sehr fehlen aus Sicht von MSD ausreichende Belege für eine generische Responderschwelle anhand eines „one-size-fits-all" Kriteriums.

Roche Pharma AG

[…] Die Vorgehensweise folgt damit das neue methodische Vorgehen des IQWiG, das in dem „Allgemeinen Methoden" 6.0 (veröffentlicht am 05.11.2020) eingeleitet wurde. Dieses Vorgehen ist, wie bereit im Stellungnahmeverfahren zum Entwurf der „Allgemeinen Methoden" Version 6.0 [65] geschildert wurde, wissenschaftlich nicht hinreichend begründet und daher aus Sicht von Roche nicht nachvollziehbar. […]

1. Bei den Responderanalysen nehmen MID („Minimal Important Differences") einen wichtigen Part ein um eine medizinisch relevante Veränderung des Zustandes von Patientinnen und Patienten zu identifizieren.

Da die MID sehr abhängig von Fragebogen und Therapiegebiet ist, kann aus Sicht von Roche eine pauschale Responseschwelle kein adäquater Ersatz für eine MID sein. Die Bestimmung der MID erfolgt heutzutage durch empirische, ankerbasierte Verfahren mittels Validierungsstudien. Dieses Verfahren sollte auch in Zukunft als Standard angesehen und anerkannt werden, vorbehaltlich einer methodischen Weiterentwicklung. Demgegenüber würden durch eine generelle Anwendung einer pauschalen Responseschwelle jegliche wissenschaftliche Erkenntnisse oder Qualitätskriterien außen vor gelassen werden.

Die vom in den "Allgemeine Methoden" 6.0 [66] vorgeschlagene pauschale 15% Responseschwelle erfolgte auf Basis unklarer Evidenz und hat einen zu generellen Anspruch in der Zusatznutzenbewertung; sie erscheint willkürlich. In den bisherigen AMNOG-Verfahren, wurden MIDs nur dann berücksichtigt, wenn sie in der wissenschaftlichen Literatur als etabliert bzw. validiert galten. Durch die Verwendung der vom IQWIG vorgeschlagenen Responseschwelle würden nun fast alle bisherigen, vom GBA akzeptierten MIDs in Frage gestellt werden. […]

AbbVie Deutschland GmbH & Co. KG

[…] aus Sicht von AbbVie [besteht] weiterhin Diskussionsbedarf hinsichtlich der Eignung der vorgeschlagenen generischen Responseschwelle von 15 %. So gibt es zunächst weder eine wissenschaftlich nachvollziehbare Rationale zur Herleitung der generischen Responseschwelle, noch gibt es eine wissenschaftliche Grundlage für einen anzunehmenden Vorteil gegenüber einer klinisch etablierten und validierten MID, die insbesondere durch ihre Definition als „minimal important difference" einen für den individuellen Patienten bedeutsamen Unterschied beschreibt und daher per se patientenrelevant ist. Des Weiteren finden patienten- und indikationsspezifische Gegebenheiten sowie Unterschiede in

63 Nordin, Å., Taft, C., Lundgren-Nilsson, Å., & Dencker, A. (2016). Minimal important differences for fatigue patient reported outcome measures—a systematic review. BMC Medical Research Methodology, 16(1), 1-16. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0167-6

64 St-Pierre, C., Desmeules, F., Dionne, C. E., Frémont, P., MacDermid, J. C., & Roy, J. S. (2016). Psychometric properties of self-reported questionnaires for the evaluation of symptoms and functional limitations in individuals with rotator cuff disorders: a systematic review. Disability and rehabilitation, 38(2), 103-122. https://www.tandfonline.com/doi/abs/10.3109/09638288.2015.1027004

65 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (2020). Dokumentation und Würdigung der Anhörung zum Entwurf der Allgemeinen Methoden 6.0. https://www.iqwig.de/methoden/allgemeine-methoden_dwa-entwurf-fuer-version-6-0_v1-0.pdf?rev=194431 [19.07.2021]

66 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (2020). Allgemeine Methoden Version 6.0. https://www.iqwig.de/methoden/allgemeine-methoden_version-6-0.pdf?rev=180500 [19.07.2021]

Skalencharakteristika bei der Verwendung einer generischen Responseschwelle keine Berücksichtigung mehr. Schließlich ist nicht hinreichend klar, wie diese vorgeschlagene generische Responseschwelle mit den bereits validierten und etablierten Responseschwellen in Form der MID im Einklang steht.

So wäre es wünschenswert, dass vor Änderung der Verfahrensordnung die potenziellen Konsequenzen der Anwendung einer generischen Responseschwelle auf die Zusatznutzenbewertung seitens des G-BA / IQWiG untersucht und transparent berichtet wird.

[…] Insgesamt weicht das Vorgehen zu Identifizierung sowie die vorgeschlagene Responseschwelle selbst vom wissenschaftlichen Vorgehen sowie dem aktuellen Stand der wissenschaftlichen Erkenntnisse ab und setzt sich ebenso über international anerkannten Kriterien und Standards der evidenzbasierten Medizin hinweg (67, 68, 69, 70, 71). Daher sollte eine Neuregelung nicht zur Ablehnung von bisher verwendeten und akzeptierten Responseschwellen bei validierten und etablierten Fragebögen / Skalen führen und es sollte insbesondere eine Konsistenz in den Bewertungskriterien innerhalb verschiedener / vergangener Nutzenbewertungen, zwischen der Nutzenbewertung und den Zulassungsverfahren (72, 73) sowie dem Vorgehen zu anderen HTA-Agenturen (74, 75, 76, 77) angestrebt werden.

UCB Pharma GmbH

Im IQWiG Methodenpapier 6.0 wird erklärt: „Eine Voraussetzung für die Berücksichtigung solcher (patientenrelevanter) Endpunkte ist die Verwendung von validierten bzw. etablierten Instrumenten“. Diese Voraussetzung sollte aus Sicht von UCB nicht nur für die Instrumente selbst, sondern auch für die Schwellenwerte für die Beurteilung der klinischen Relevanz gelten. Das IQWiG referenziert im Methodenpapier 6.0 insgesamt 8 systematisch

67 Carrasco-Labra A, Devji T, Qasim A, Phillips MR, Wang Y, Johnston BC, et al. Minimal important difference estimates for patient-reported outcomes: A systematic survey. J Clin Epidemiol. 2021;133:61-71.
68 Coens C, Pe M, Dueck AC, Sloan J, Basch E, Calvert M, et al. International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium. Lancet Oncol. 2020;21(2):e83-e96.
69 Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ. 2020;369:m1714.
70 Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147-57.
71 Reni M, Braverman J, Hendifar A, Li CP, Macarulla T, Oh DY, et al. Evaluation of Minimal Important Difference and Responder Definition in the EORTC QLQ-PAN26 Module for Assessing Health-Related Quality of Life in Patients with Surgically Resected Pancreatic Adenocarcinoma. Ann Surg Oncol. 2021.
72 European Medicines Agency. REFLECTION PAPER ON THE REGULATORY GUIDANCE FOR THE USE OF HEALTHRELATED QUALITY OF LIFE (HRQL) MEASURES IN THE EVALUATION OF MEDICINAL PRODUCTS. 2005.https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-regulatory-guidance-use-healthrelated-quality-life-hrql-measures-evaluation_en.pdf
73 European Medicines Agency. Appendix 2 to the guideline on the evaluation of anticancer medicinal products in man. The use of patient-reported outcome (PRO) measures in oncology studies. 2016.https://www.ema.europa.eu/en/documents/other/appendix-2-guideline-evaluation-anticancer-medicinal-products-man_en.pdf
74 Brazier J LL. NICE DSU TECHNICAL SUPPORT DOCUMENT 8: AN INTRODUCTION TO THE MEASUREMENT AND VALUATION OF HEALTH FOR NICE SUBMISSIONS. 2011.https://www.ncbi.nlm.nih.gov/books/NBK425820/pdf/Bookshelf_NBK425820.pdf
75 Canadian Agency for Drugs and Technologies in Health. Guidelines for the Economic Evaluation of Health Technologies: Canada 4th Edition. 2017.https://cadth.ca/node/101497
76 European Network For Health Technology Assessment. Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints. 2015.https://www.eunethta.eu/wp-content/uploads/2018/02/WP7-SG3-GL-clin_endpoints_amend2015.pdf
77 Haute Autorité de Santé. Evaluation of Health Technologies at HAS: Role of Quality of Life. 2018. https://www.has-sante.fr/jcms/c_2883073/fr/evaluation-des-technologies-de-sante-a-la-has-place-de-la-qualite-de-vie

recherchierte systematische Übersichtsarbeiten zu Minimal Important Differences (MIDs). Aus diesen „wurde ein Wert von 15 % der Spannweite der jeweiligen Skalen als plausibler Schwellenwert für eine eher kleine, aber hinreichend sicher spürbare Veränderung identifiziert“. Eine Begründung für die Plausibilität des Schwellenwertes 15 % der Skalenspannweite wird im Methodenpapier 6.0 nicht geliefert. Es wird lediglich davon ausgegangen, dass die 15 %-Schwelle eine hinreichend spürbare Veränderung darstellt. Das heißt, es handelt sich bei dieser Schwelle nicht mehr um eine MID, sondern um eine Responseschwelle, die in den meisten Fällen größer oder gleich publizierter MIDs ist. Darin liegt aber genau ein Problem dieser fixen Responseschwelle. Es handelt sich nicht mehr um eine patientenrelevante Responseschwelle, sondern um einen Schwellenwert, der unabhängig von der Indikation und Patientenpopulation für alle Instrumente angewendet wird.

Des Weiteren sind die Eigenschaften dieser fixen Responseschwelle vor der Einführung im Methodenpapier 6.0 nur auf Basis systematischer Literaturrecherchen hergeleitet worden. Wie eine Synthese aus den Resultaten dieser verschiedenen identifizierten Quellen methodisch durchgeführt wurde bleibt unklar. […]

Die Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen im Rahmen der Dossiererstellung ist mit der Einführung der vom IQWiG im Methodenpapier 6.0 beschriebenen Methodik aufgrund von einer nicht nachtvollziehbaren Verallgemeinerung über alle Indikationen hinweg, der Vernachlässigung der Patientenperspektive bzgl. der Definition „für den Patienten hinreichend sicher spürbare Änderung“ und eines Verlustes in der Trennschärfe in vielen Situationen nicht mehr sinnvoll. Darüber hinaus ist eine solche Schwelle nicht länger klinisch relevant. Daher ist ein solches Vorgehen aus Sicht von UCB abzulehnen.

GlaxoSmithKline GmbH & Co. KG

[…] Die Änderungen stoßen bei GSK auf große Bedenken, die bereits im IQWiG-Stellungnahmeverfahren adressiert wurden. Diese Bedenken werden im Folgenden dargestellt und begründet.

Fehlende Wissenschaftliche Validität

Das vorgeschlagene Vorgehen, ein pauschales Responsekriterium von 15 % der Skalenspannweite für alle Erhebungsinstrumente und für alle Indikationen anzuwenden, ist aus unserer Sicht auch weiterhin wissenschaftlich nicht valide begründet und daher für uns nicht nachvollziehbar. Zur Herleitung dieser Schwelle bezieht sich das IQWiG in seinem Methodenpapier zwar auf eine „Sichtung von systematisch recherchierten systematischen Übersichtsarbeiten zu MIDs“, es bleibt aber immer noch unklar, wie dieser Schwellenwert konkret hergeleitet wird. Darauf hatte GSK, neben vielen anderen Stellungnehmern, auch schon im Stellungnahmeprozess zum IQWiG-Methodenpapier 6.0 hingewiesen (78). Auch unter Berücksichtigung der vom IQWiG vorgenommenen Würdigung der zahlreichen Stellungnahmen im o.g. Stellungnahmeverfahren zu diesem Punkt liegt keine wissenschaftlich belastbare Rationale für den gewählten Schwellenwert vor (79).

Es handelt sich weiterhin um einen solitären IQWiG-Vorschlag, der nicht als ein in der Wissenschaft akzeptiertes Kriterium der evidenzbasierten Medizin angesehen werden kann. Dazu fehlt insbesondere eine in einem wissenschaftlichen, Peer-reviewed Journal

78 GSK, GlaxoSmithKline. Stellungnahme zum Entwurf der Allgemeinen Methoden Version 6.0 2019 31.01.2020. Available from: https://www.iqwig.de/methoden/allgemeine-methoden_dwa-entwurffuer-version-6-0_v1-0.pdf?rev=194431.

79 IQWiG, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Dokumentation und Würdigung der Anhörung zum Entwurf der Allgemeinen Methoden 6.0 2020 28.06.2021. Available from: https://www.iqwig.de/methoden/allgemeine-methoden_dwa-entwurf-fuer-version-6-0_v1-0.pdf?rev=194431.

veröffentlichte Arbeit zu diesem Ansatz, die eine Basis für eine breite, internationale wissenschaftliche Diskussion darstellen würde.

In der wissenschaftlichen Literatur wurde dazu kürzlich ein Instrument vorgestellt, das die Zuverlässigkeit von MIDs bestimmt. Dazu wurde ein Kriterienkatalog aufgestellt, bei dem u.a. auch ankerbasierte Verfahren und eine ausreichende Korrelation zwischen Anker und PRO eingehen. (80, 81). Dieses Instrument ist aktuell Gegenstand einer wissenschaftlichen Diskussion. So haben andere Forschungsgruppen Nachbesserungen hinsichtlich der notwendigen Korrelation gefordert (82). Diese wichtige wissenschaftliche Diskussion ist aber noch nicht abgeschlossen. Sie findet, anders als der IQWiG-Vorschlag, jedoch in der weltweiten wissenschaftlichen Gemeinschaft insbesondere in Form von Veröffentlichungen in wissenschaftlichen Journalen statt.

Neben der Willkürlichkeit dieser Setzung bezweifeln wir die Sinnhaftigkeit eines pauschalen Schwellenwertes für alle Erhebungsinstrumente und für alle Indikationen. […]

Die Autoren (83) kommen in ihrer Arbeit daher zu folgender Empfehlung „We recommend that the MID is based primarily on relevant patient-based and clinical anchors, with clinical trial experience used to further inform understanding of MID."

Dieser Empfehlung schließen wir uns an, da nur durch die Berücksichtigung von Patienten- und Arztbasierten Ankern sichergestellt werden kann, dass die abgeleitete MID auch klinisch relevant und für den Patienten spürbar ist. Diese Sichtweise wird unter anderem auch von der EMA geteilt (84).

Auch aktuelle Arbeiten zur Herleitung von MIDs in spezifischen Indikationen basieren im Wesentlichen auf ankerbasierten Verfahren und benutzen weitere Methoden zur Triangulation (z.B. Sully, et al.,2019 (85)) für die Indikation Multiples Myelom).

Bristol-Myers Squibb GmbH & Co. KGaA

BMS begrüßt, dass sich IQWiG und G-BA mit der adäquaten Analyse des komplexen Konstrukts der Lebensqualitätsfragebögen auseinandersetzen. Auch ist eine Verfahrenssicherheit in Bezug auf die Akzeptanz der Analysen, insbesondere wenn es sich um Analysen unter Verwendung einer Responseschwelle handelt, wünschenswert.

Der aktuell vorgeschlagene Ansatz, Responderanalysen nur noch dann zu akzeptieren, wenn sie präspezifiziert sind und >15% der Skalenspannweite betragen, oder wenn sie post hoc mit genau 15% der Skalenspannweite durchgeführt wurden, ist jedoch aus Ansicht von BMS nicht sachgerecht. Dieser Ansatz basiert auf einer nicht systematisch durchgeführten Literaturrecherche des IQWiG (86) und lässt jegliche etablierten und wissenschaftlich

80 Devji T; Carrasco-Labra A; Qasim A; Phillips M; Johnston BC; Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. Bmj. 2020; 369.

81 Carrasco-Labra A; Devji T; Qasim A; Phillips MR; Wang Y; Johnston BC, et al. Minimal important difference estimates for patient-reported outcomes: A systematic survey. Journal of Clinical Epidemiology. 2020.

82 Peipert J. Rapid Response: Re: Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ. 2020; 369.

83 Revicki D; Hays RD; Cella D; Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of clinical epidemiology. 2008; 61(2): 102-9.

84 EMA, European Medicines Agency. Appendix 2 to the guideline on the evaluation of anticancer medicinal products in man The use of patient-reported outcome (PRO) measures in oncology studies 2016 13.01.2020. Available from: https://www.ema.europa.eu/en/documents/other/appendix-2-guideline-evaluation-anticancer-medicinal-products-man_en.pdf.

85 Sully K; Trigg A; Bonner N; Moreno-Koehler A; Trennery C; Shah N, et al. Estimation of minimally important differences and responder definitions for EORTC QLQ-MY20 scores in multiple myeloma patients. European journal of haematology. 2019; 103(5): 500-9.

86 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (2021). Allgemeine Methoden Version 6. https://www.iqwig.de/methoden/allgemeine-methoden_version-6-0.pdf?rev=180500

untersuchten Responseschwellen unberücksichtigt. Eine systematische Untersuchung der Validität der Schwelle und deren Einfluss auf die Nutzenbewertung ist bislang nicht erfolgt. Zudem steht der Vorschlag im Widerspruch zur Bewertungspraxis bei den Zulassungs- und europäischen HTA-Behörden (87, 88). Diese orientieren sich vielmehr an etablierten MIDs (wie sie z.B. auch die Cochrane Collaboration (89) verwendet) und etablierten Qualitätskriterien (etwa anker-basierten Verfahren, etc.) und befürworten ausdrücklich kein generisches Responsekriterium über alle Instrumente bzw. Therapiesituationen hinweg. […]

Merck Serono GmbH

[…] Diese Änderungen stoßen seitens Merck auf erhebliche Bedenken, die bereits im Stellungnahmeverfahren zum Entwurf der „Allgemeinen Methoden“ Version 6.0 geschildert wurden. […]

Die Anwendung einer generischen, universellen anzuwendenden MID-Schwelle von 15 % der theoretisch möglichen Skalenspannweite über Erhebungsinstrumente der verschiedensten Indikationen hinweg, ist wissenschaftlich nicht ausreichend begründet und kann daher nicht unterstützt werden. Die Herleitung dieser These („eine Sichtung von systematischen Übersichtsarbeiten“) wird im IQWiG-Methodenpapier 6.0.nicht dargestellt und bleibt daher nicht nachvollziehbar. Die vom IQWiG zitierten Referenzen (90, 91) thematisieren einzelne Symptome wie „fatigue“ und „symptoms and functional limitations in individuals with rotator cuff disorders“. Eine Übertragbarkeit auf krankheitsspezifische wie generische Erhebungsinstrumente bleibt fraglich. […]

D.h. Veränderungen sind unterschiedlich spürbar, abhängig von der Erkrankung und der Richtung der Veränderung. Ein universeller Schwellenwert kann dies nicht adäquat berücksichtigen.

Die Evidenz der Aussage, dass „sich MIDs in vielen Fällen zwischen 10 % und 20 % der Spannweite der Scores eines Erhebungsinstruments bewegen“, bleibt unerklärt. Es ist nicht nachzuvollziehen, welche Kriterien bei der Recherche zugrunde lagen und was tatsächlich unter „vielen Fällen“ zu verstehen ist. Auf Basis dieser nicht näher bezeichneten Evidenz soll eine Schwelle für alle MID in Höhe von 15 % abgeleitet werden. Ob die Wahl auf den Mittelwert einer konkreten Begründung folgt, bleibt unklar. Eine vfa-Überprüfung aller bisher im AMNOG-Verfahren akzeptierten MID zeichnet dagegen ein anderes Bild. Hier liegen die MID-Spannweiten zwischen 2 % und 40 % und erreichen im Mittel ca. 9 %. Der mittlere Wert ist niedriger als angenommen.

Daher sollte ein universeller Schwellenwert der MID von 15 % keine Berücksichtigung in den neuen Modulvorlagen finden. Vielmehr sollten weiterhin etablierte Schwellenwerte mit bekannten psychometrischen Eigenschaften anerkannt werden.

87 Chassany O. Is “15% of the scale range” Universally Applicable to Define “MID” and Clinical Relevance of Patient-Reported Treatment Benefits? Group Discussion with the ISPOR Clinical Outcome Assessment Special Interest Group (COA SIG). Virtual ISPOR 2021, Group Discussion (GD3).

88 Shaw J. Industry perspective. Is “15% of the scale range” Universally Applicable to Define “MID” and Clinical Relevance of Patient-Reported Treatment Benefits? Group Discussion with the ISPOR Clinical Outcome Assessment Special Interest Group (COA SIG). Virtual ISPOR 2021, Group Discussion (GD3).

89 Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 6.2, 2021. Chapter 15: Interpreting results and drawing conclusions. 15.5.3.5 Presenting continuous results as minimally important difference units. www.training.cochrane.org/handbook

90 Nordin et al - MID for fatigue patient reported outcome measures - a systematic review

91 St-Pierre 2016

Ungeachtet der Herleitung der Schwelle für MID ist es aus Sicht von Merck zu berücksichtigen, dass die Validierung einer MID keineswegs pauschalisiert werden kann. Aus der Arbeit von Revicki et al. 5 (92): […]

„We recommend that the MID is based primarily on relevant patient-based and clinical anchors, with clinical trial experience used to further inform understanding of MID." (92)

Merck befürwortet diese Empfehlung, patientenberichtete Anker in der Validierung zu nutzen, um die klinische Relevanz einer bestimmten MID zu gewährleisten. Diese Sichtweise wird unter anderem auch von den Zulassungsbehörden geteilt6.

Biogen GmbH

[…] Biogen begrüßt, dass Standards zur Bewertung von Studienergebnissen hinsichtlich ihrer klinischen Relevanz für Patienten stetig weiterentwickelt werden. Aus Sicht von Biogen ist das in der aktuellen Änderung gewählte Vorgehen jedoch sowohl auf Basis der aktuell zur Verfügung stehenden wissenschaftlichen Evidenz als auch aus klinischer Sicht problematisch.

Biogen erkennt an, dass für einzelne Instrumente unterschiedliche (ankerbasierte) individuelle MIDs (Minimal Important Difference) mit z. T. großen Spannweiten vorliegen können. Diese Variabilität spiegelt nach Ansicht von Biogen allerdings weniger eine Unsicherheit hinsichtlich der Interpretation berichteter Ergebnisse bzw. der Ableitung einer klinischen Relevanz wider, vielmehr werden hierdurch die für die jeweilige Untersuchungspopulation spezifischen und damit bewertungsrelevanten Eigenschaften abgebildet. In diesem Zusammenhang ist aus unterschiedlichen Erkrankungen bekannt, dass sich die MID für Patienten mit unterschiedlichem Schweregrad der Erkrankung für dasselbe eingesetzte Instrument signifikant unterscheidet (z.B. Anstieg der MID zur Bestimmung klinisch relevanter Verschlechterung mit zunehmender Schwere der Erkrankung). Es kann daher angenommen werden, dass einzig die Bestimmung einer patientenindividuellen bzw. populationsspezifischen MID – gegenüber einer „fixen" prädefinierten Schwelle – die Patientenperspektive adäquat abbildet.

Aus Sicht von Biogen lässt der Ansatz einer „one-size-fits-all"-Lösung, d. h. eines fixen, prädefinierten Schwellenwerts, angewendet auf alle skalenbasierten Untersuchungsinstrumente und unabhängig von der spezifischen Indikation und untersuchten Population, die individuelle Wahrnehmung des Patienten vollständig außer Acht. Sie kann somit aus Sicht von Biogen keine adäquate Lösung darstellen, um die klinische Relevanz von Ergebnissen mit minimaler Unsicherheit abzubilden. […]

Insbesondere vor dem Hintergrund, dass sich der Nutzen eines Arzneimittels direkt aus dem patientenrelevanten therapeutischen Effekt (u. a. hinsichtlich der häufig auf Basis von skalenbasierten Instrumenten erhobenen Verbesserung des Gesundheitszustands oder der Lebensqualität) ableitet2, ist die Einführung einer solchen indikations- und populations-unabhängigen Relevanzschwelle kritisch zu bewerten.

Deutsche Diabetes Gesellschaft e. V.

Die DDG sieht in einer Konkretisierung zur Ergebnisdarstellung von PRO einen Fortschritt, stellt jedoch die empirische Grundlage für die prioritäre Festlegung der MID auf 15% der Skalenspannweite in Frage, die vom IQWiG im Methodenpapier 6.0 nicht hinreichend belegt ist. In diesem Zusammenhang weist die DDG – wie auch der G-BA in seinem Entwurf für die Tragenden Gründe im Text "2. Eckpunkte der Entscheidung" - darauf hin, dass neben anderen Faktoren die Schwere einer Erkrankung und das Instrument (Fragebogen) zur Erfassung der PROs selbst einen Einfluss auf die MID haben. Aus Sicht der DDG benötigt daher z.B. eine kränkere Stichprobe einen anderen Wert der Skalenspannweite als eine gesündere Population

92 Revicki et al - Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes

In diesem Zusammenhang sind außerdem in Abhängigkeit der Studienpopulation und der Studiendesigns "Ceiling-Effekte" und "Bodeneffekte" zu berücksichtigen, die auch nicht vollständig durch den Umgang mit Responsekriterien wie in Punkt 2 oben beschrieben, behandelt werden können. Aus Sicht der DDG sollten möglichst Fragebögen verwendet werden, die in Bezug auf die Erkrankung und einem MID bereits evaluiert sind. In diesem Falle sollten für den jeweiligen Fragebogen und Krankheitsbild evaluierte MID-Schwellen verwendet werden. […]

Pfizer Deutschland GmbH

[…] Bisher wurden in der Nutzenbewertung nur jene MID berücksichtigt, die in der wissenschaftlichen Literatur als etabliert bzw. in der jeweiligen Indikation als validiert galten. Das neu vorgeschlagene Vorgehen ermöglicht eine Vorhersehbarkeit der Akzeptanz von Responseschwellen und bietet zudem einen Weg, relevante Änderungen für Erhebungsinstrumente nachzuweisen, für die keine validierte MID in der entsprechenden Indikation vorliegt.

Nach Ansicht Pfizers sollten validierte, allgemein akzeptierte MIDs jedoch weiterhin Verwendung als Responsekriterien in den Bewertungsverfahren finden. Selbst wenn bezüglich der Qualität der Validierungsstudie der MID aufgrund einzelner Kriterien Unsicherheiten bestehen, sollte eine Diskussion über die Aussagekraft der ermittelten MID ermöglicht und nicht von vornherein ausgeschlossen werden. Eine inhaltliche Bewertung der MIDs und ihrer Validierungsstudien findet gemäß des neu geplanten Vorgehens nicht mehr statt.

Um eventuell bestehende Unsicherheiten hinsichtlich der Qualität von Validierungsstudien zu MIDs beseitigen zu können, sollten allgemein Kriterien für die Beurteilung der Validierungsstudien formuliert werden. Erste Ergebnisse werden bereits in der wissenschaftlichen Literatur diskutiert (93, 94, 95). Die Bestimmung eines Schwellenwerts für die klinische Relevanz einer Veränderung sollte generell mit Hilfe statistischer Methoden gerechtfertigt werden (z.B. durch ein ankerbasiertes Verfahren). Auch wenn für ein Instrument mehrere validierte MIDs vorliegen, sollten nicht alle MIDs kategorisch abgelehnt werden. In dieser Situation sollte das Risiko einer möglichen ergebnisgesteuerten Berichterstattung durch eine beliebige Auswahl einer MID vielmehr dadurch minimiert werden, dass eine konkrete Methode genannt wird, wie das Responsekriterium basierend auf den validierten MIDs zu wählen ist. Beispielsweise könnte die MID anhand der Ähnlichkeit zur untersuchten Indikation gewählt werden und Sensitivitätsanalysen mit weiteren publizierten MIDs durchgeführt werden.

Der nun gewählte generische Schwellenwert in Höhe von 15% der Skalenspannweite ist aus wissenschaftlicher Sicht problematisch. Die Festlegung des 15%-Schwellenwertes durch das IQWiG wird kritisch gesehen, da sie arbiträr auf der Sichtung von wenigen ausgewählten systematischen Übersichtsarbeiten zu MID in wenigen Indikationen beruht (96). Es ist sehr fraglich, ob die Ergebnisse dieser Analyse auf alle Erhebungsinstrumente und Indikationen übertragen werden können. Der gewählte Schwellenwert in Höhe von 15% der Skalenspannweite würde bei vielen etablierten Instrumenten, für die akzeptierte validierte MIDs vorliegen, zu deutlich höheren Responseschwellen führen (97). […]

93 Coon C, Cook K. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores Quality of Life Research 2018;27(1):33-40.

94 Musoro ZJ, Hamel J-F, Ediebah DE, Cocks K, King MT, Groenvold M, et al. Establishing anchor-based minimally important differences (MID) with the EORTC quality-of-life measures: a meta-analysis protocol. BMJ Open. 2018;8(1):e019117.

95 U.S. Department of Health and Human Services Food and Drug Administration. Guidance for Industry, Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims: draft guidance Health and Quality of Life Outcomes 2006;4.

96 IQWIG. Allgemeine Methoden Version 6.0; 2020 URL: www.iqwig.de.

97 vfa. Stellungnahme zum Entwurf der IQWiG "Allgemeinen Methoden" Version 6.02020.

Abgesehen von den oben geschilderten Argumenten sollten für eine weiterführende Diskussion auch folgende Punkte berücksichtigt werden:

- Während das IQWiG im Entwurf für Version 6.0 vom 05.12.2019 nur zwei Literaturstellen angegeben hatte, um den Schwellenwert von 15% abzuleiten (98) (wahrscheinlich als Mittelwert zwischen 10% und 20%), sind es in der finalen Version acht Literaturstellen – am Schwellenwert hat sich jedoch nichts geändert (1). Das IQWiG fordert in seinen Allgemeinen Methoden Version 6.0 (Kapitel 9.3.3): „Voraussetzung für die Berücksichtigung solcher Endpunkte ist die Verwendung von validierten bzw. etablierten Instrumenten“ - die Ableitung des 15%-Schwellenwertes durch das IQWiG erscheint dagegen nicht validiert bzw. etabliert und somit nicht wissenschaftlich nachvollziehbar. […]
- Anstatt sich auf ein einziges (sicherlich für Firmen und IQWiG/ G-BA) einfach anzuwendendes, aber wissenschaftlich höchst umstrittenes Responsekriterium zu fixieren, sollte mit erfahrenen Wissenschaftler*Innen und auch Patientenvertreter*Innen ein Kriterienkatalog aufgestellt und auch eine breiter angelegte Diskussion bezüglich anderer Auswerteverfahren (zum Beispiel zeitabhängige Verfahren, Analyse der Fläche unter der Kurve etc.) erfolgen. Mit Verabschiedung eines 15% „one-size-fits-all“ Kriteriums wird eine inhaltliche Weiterentwicklung zur Auswertung dieser patientenrelevanten Endpunkte gehemmt werden.

Collegium Internationale Psychiatriae Scalarum (CIPS)

[…] Die einfache Lösung einer komplexen, multifaktoriellen Problematik, wie im IQWiG-Methodenpapier detailliert erörtert, erinnert an die Lösung des gordischen Knotens durch Durchschlagen, kann aber nicht überzeugen.

Kritikpunkte

Nummerischer Wert der Responseschwelle von 15% der Skalenspannweite

Dieser Schwellenwert wird im IQWiG-Methodenpapier als „plausibel“ bezeichnet, ist aber letztlich willkürlich und kann, bei mangelnder empirischer Fundierung, je nach Interessenlage jederzeit geändert werden. Es könnten mit unveränderter Argumentation z.B. auch 10% oder 20% der Skalenspannweite (oder jeder andere Wert) vereinbart werden.

Generische Anwendung des gleichen Schwellenwerts

Durch die einheitliche prozentuale Definition wird zwar eine Standardisierung der MCID in Bezug auf die Skalenspannweite erreicht, dabei aber durch unterschiedliche Spannweiten bedingte Unterschiede in der Differenzierungsfähigkeit der Skalen vernachlässigt.

Weiterhin bleiben bei der uniformen Anwendung der 15%-Schwelle skalenunabhängige Faktoren, wie die Patientenpopulation, der Schweregrad der Symptomatik, das Therapieziel oder der situative Kontext unberücksichtigt. […]

Schlussfolgerung

Die Definition von klinisch relevanten Differenzen von Skalenwerten auf Individualebene setzt die Kenntnis der Differenzierungsfähigkeit der Skala im Sinne von minimalen, zuverlässigen Differenzen voraus. Aus Sicht von CIPS sollte eine MCID skalenspezifisch, unter Berücksichtigung psychometrischer Eigenschaften, und nur im Kontext der Bedingungen der jeweiligen klinischen Studie definiert werden. Besondere Merkmale der klinischen Studie sind dabei das Patientenkollektiv, Schweregrad und Dauer der Erkrankung, begründete Erwartungen an Effekte und Nebenwirkungen.

98 IQWiG. Entwurf Allgemeine Methoden 6.0 2019.

Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V.

[…] Die folgende Stellungnahme wurde von der ständigen Kommission Nutzenbewertung der AWMF verfasst. Die Deutsche Gesellschaft für Rheumatologie (DGRh), die Deutsche Gesellschaft für Pneumologie und Beatmungsmedizin (DGP) sowie die Deutsche Gesellschaft für Allergologie und klinische Immunologie (DGAKI) schließen sich der Stellungnahme an. […]

- Die Weiterentwicklung der Nutzenbewertung nach § 35a SGB V im Bereich der Ergebnisdarstellung von patientenberichteten Endpunkten (PRO) wird begrüßt. […]
- Eine Vereinheitlichung der Bewertungskriterien für Responder-Analysen in den z. T. komplexen Skalen wird ebenfalls begrüßt.
- Die Festlegung eines Grenzwertes von 15% wird hinterfragt. Der Entwurf greift einen Vorschlag des IQWiG auf, ist aber in dem Methodenpapier 6.0 nicht hinreichend belegt und im strengen Sinne nicht evidenz-basiert. […]

Deutsche Gesellschaft für Innere Medizin e. V.

[…] Die Deutsche Gesellschaft für Innere Medizin (DGIM) verlässt sich hier auf die Methodenrecherche des IQWiG, möchte aber dennoch einen wichtigen Punkt zu bedenken geben: Neben der Quantifizierung solcher Effekt erscheinen uns die Parameter, die abgefragt werden, wichtig (er). Diese sollten relevant für die jeweilige Krankheitsentität sein, aber auch für das Stadium der Erkrankung. D.h. welche(r) Parameter wird/werden herangezogen? Eine Veränderung von 15 % (Schwelle) des Erhebungskriteriums in einem frühen/leichten Stadium einer Erkrankung muss möglicherweise ganz anders bewertet werden als in einem sehr späten/schweren Stadium einer Erkrankung. Hinsichtlich dieser inhaltlichen (nicht methodischen Fragen) erscheint es uns wichtig, die geeigneten Vertreter der Patienten und mit den jeweiligen Krankheitsbildern vertrauten Ärzte heranzuziehen.

IQVIA Commercial GmbH & Co. OHG

Mit den Änderungen der Modulvorlage übernimmt der G-BA unverändert den Vorschlag des IQWiG aus dem Methodenpapier 6.0. Die von IQVIA im Stellungnahmeverfahren zum Methodenentwurf des IQWiG vorgebrachten Kritikpunkte (IQVIA Commercial GmbH & Co. OHG 2019) haben nach wie vor Bestand. Die geäußerten Bedenken konnten seitens des IQWiG weder im Rahmen der Diskussion zur wissenschaftlichen Erörterung der Stellungnahmen zum Entwurf des Methodenpapiers noch durch ergänzenden Erläuterungen in der finalen Fassung des Methodenpapiers aus dem Weg geräumt werden.

Diese sind:

- das Fehlen einer wissenschaftlich fundierten Herleitung des Schwellenwerts – auch die in der finalen Version des Methodenpapiers als Quellen angegebenen Literaturstellen lassen weder auf eine systematische Identifikation noch auf eine Bewertung der Aussagesicherheit der identifizierten Studien schließen
- die fehlende Diskussion und Berücksichtigung wissenschaftlicher Empfehlungen, neuer wissenschaftlicher Bemühungen; aus unserer Sicht ist dies nicht zu vereinbaren mit dem Grundsatz der evidenzbasieren Medizin
- die Wahl eines singulären Kriteriums, das keine differenzierte Betrachtung zulässt; sowohl in Bezug auf Indikation, die Spezifika der Skala, generisches oder krankheitsspezifisches Instrument, die Richtung der untersuchten Veränderung (Verbesserung / Verschlechterung), etc.
- die Operationalisierung eines Responsekriteriums anhand eines Prozentwertes der Skalenspannweite ist möglicherweise eine schlechte Wahl, da sie die Verteilung der Scorewerte außer Acht lässt; von einer gleichmäßigen Verteilung der Scorewerte ausgeht

- die fehlende Gleichbewertung zwischen einzelnen Nutzenbewertungsverfahren

Verband Forschender Arzneimittelhersteller e. V.

[…] Trotz der anhaltenden Diskussion um die Verbesserung der Standards hat das IQWiG überraschend einen eigenen Bewertungsansatz für Responderschwellen entwickelt, der alternativ zu MID verwendet werden soll. Dieser Ansatz muss grundsätzlich hinterfragt werden. Er erscheint aus Sicht des vfa als übereilt und wissenschaftlich nicht hinreichend begründet. Er verlässt zum einen die Ausrichtung an den seitens des IQWiG selbst in den letzten Jahren der Bewertungspraxis vorgebrachten möglichen Standards und lässt den internationalen Entwicklungsansatz gänzlich außer Acht [99, 100, 101]. Der IQWiG-Vorschlag steht zugleich im Widerspruch zu aktuellen Vorgaben der Zulassungsbehörden sowie anderer HTA-Organisationen [102, 103, 104]. Diese orientieren sich vielmehr an etablierten MID (wie z. B. die Cochrane Collaboration) und zuvor genannten Qualitätskriterien (wie z. B. ankerbasierte Verfahren, etc.) und befürworten ausdrücklich kein generisches Responsekriterium für alle Instrumente bzw. Therapiesituationen.

Dessen ungeachtet soll durch die Festsetzung eines eigenen Richtmaßes in einem „one-size-fits-all"-Ansatz des IQWiG eine strengere MID-Bewertung in „pragmatischer Weise" bewerkstelligt werden. Das Richtmaß in Form eines 15-%-Schwellenwerts soll dabei eine schnelle Beurteilung der „Mindestgröße" einer geeigneten MID erlauben.

Gleichzeitig wird die Eignung fast aller bisherigen MID für die Nutzenbewertung in Frage gestellt und die Notwendigkeit einer strengeren Handhabung betont. Eine inhaltliche Bewertung der MID bzw. ihrer Validierungsstudien oder eine Diskussion über die Aussagekraft der MID (z. B. mit Hilfe von Sensitivitätsanalysen) findet nach dieser Vorgehensweise gar nicht statt. Anstatt möglicher Qualitätskriterien soll hierfür ein generisches Richtmaß zur Anwendung kommen. Diese Vorgehensweise ist somit weder allgemein akzeptiert noch etabliert. Sie entspricht nicht dem aktuellen Stand der wissenschaftlichen Erkenntnisse sowie den international anerkannten Kriterien und Standards der evidenzbasierten Medizin.

Empirische Analyse aller MIDs in der Nutzenbewertung

Der „one-size-fits-all"-Ansatz eines generischen Richtmaßes ist aus wissenschaftlicher Sicht problematisch. Dabei ist bereits die Festsetzung des 15-%-Schwellenwerts zu hinterfragen. Dieser wurde nicht anhand von nachvollziehbaren Kriterien und unter Beteiligung von Patientinnen und Patienten ermittelt, sondern folgt einer arbiträren Festlegung anhand der

99 Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, Zeraatkar D, Bhatt M, Jin X, Brignardello-Petersen R, Urquhart O, Foroutan F, Schan-delmaier S, Pardo-Hernandez H, Vernooij RW, Huang H, Rizwan Y, Siemieniuk R, Lytvyn L, Patrick DL, Ebrahim S, Furukawa T, Nesrallah G, Schünemann HJ, Bhandari M, Thabane L, Guyatt GH. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ 2020 Jun 4;369:m1714. doi: 10.1136/bmj.m1714 https://www.bmj.com/content/ 369/bmj.m1714

100 Peipert JD, Cella D. Rapid Response to "Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study" (doi: https://doi.org/10.1136/bmj.m1714) 06. September 2020 https://www.bmj.com/content/369/bmj.m1714/rapid-responses

101 IQWiG. Cabozantinib (Nierenzellkarzinom) – Addendum zum Auftrag A17-56 (Addendum A18-13 vom 09.03.2018

102 Chassany O. Is "15% of the scale range" Universally Applicable to Define "MID" and Clinical Relevance of Patient-Reported Treatment Benefits? Group Discussion with the ISPOR Clinical Outcome Assessment Special Interest Group (COA SIG). Virtual ISPOR 2021, Group Discussion (GD3).

103 Shaw J. Industry perspective. Is "15% of the scale range" Universally Applicable to Define "MID" and Clinical Relevance of Patient-Reported Treatment Benefits? Group Discussion with the ISPOR Clinical Outcome Assessment Special Interest Group (COA SIG). Virtual ISPOR 2021, Group Discussion (GD3).

104 Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 6.2, 2021. Chapter 15: Interpreting results and drawing conclusions. 15.5.3.5 Presenting continuous results as minimally important difference units. www.training.cochrane.org/handbook

vom IQWiG in einigen ausgewählten Literaturquellen beobachteten MID-Spannweiten für wenige Indikationen. Die Festlegung des 15-%-Schwellenwerts erfolgte im Entwurf des Methodenpapiers zunächst als simpler Mittelwert zwischen 10 % und 20 % der Spannweite der Scores in zwei Publikationen. In der Dokumentation und Würdigung der Anhörung gibt das IQWiG anschließend zwar eine deutlich größere Bandbreite der MID an (1 % bis 38 %), verbleibt jedoch bei seiner Festsetzung [105].

Eine vfa-Überprüfung aller bisher in AMNOG-Verfahren akzeptierten MID zeigt dabei, dass die MID-Spannweiten hier zwischen 2 % und 40 % liegen und im Mittel ca. 9 % erreichen [106]. Der mittlere Wert ist somit deutlich niedriger als vom IQWiG angenommen. Dies zeigt deutlich, wie individuell verschieden einzelne MID für Patientinnen und Patienten mit einer bestimmten Erkrankung in speziellen Fragebögen sein können und müssen. Der festgelegte „one-size-fits-all"-Ansatz kann bekannte Unterschiede der Patientensicht auf bedeutsame Ergebnisse nicht hinreichend berücksichtigen und ist folglich nicht für alle Therapiesituationen gleichermaßen sinnvoll.

Eine Analyse zum Vergleich des 15-%-Schwellenwerts mit den in AMNOG-Verfahren bislang verwendeten MID-Schwellen wurde seitens des IQWiG nicht durchgeführt. Es zeigt sich jedoch, dass das festgesetzte Richtmaß regelhaft höher ist als die MID-Schwellen, die bislang in AMNOG-Verfahren akzeptiert wurden. Eine umfassende vfa-Überprüfung aller bisher im AMNOG-Verfahren akzeptierten MID zeigt, dass das Richtmaß in fast 90 % der Fälle teils deutlich (um das 1,1-fache bis 8,3-fache) höher liegt [106]. Durch die Anwendung des Richtmaßes würden damit fast alle bisher im AMNOG akzeptierten MID Ihre Gültigkeit verlieren. International etablierte patientenberichtete Auswertungen wären damit ungeeignet, während die Messlatte für den Nachweis von therapeutischen Verbesserungen oder Verschlechterungen substanziell höher wäre. Die teils deutlichen Unterschiede offenbaren dabei die individuellen, patientengerechten MID Anforderungen, welche durch einen generischen Schwellenwert nicht hinreichend berücksichtigt werden können.

Janssen-Cilag GmbH

In den bis zum Inkrafttreten des IQWiG-Methodenpapiers 6.0 erfolgten Nutzenbewertungsverfahren wurden sowohl vom IQWiG als auch vom G-BA ausnahmslos nur Responsekriterien akzeptiert, die nach wissenschaftlichen Methoden als validiert gelten oder die ihrerseits etabliert sind, um den Nachweis eines klinisch relevanten Effekts vorzunehmen (107).

Validierungen von MCID werden für gewöhnlich im Rahmen von Studien ermittelt. An diese Validierungsstudien werden hohe Maßstäbe angelegt. Es werden ankerbasierte Verfahren vor distributionsbasierten Verfahren bevorzugt und es sind Längsschnittstudien anzuwenden (108). […] Die MCID für ein Erhebungsinstrument ist somit abhängig von dem jeweiligen Instrument und der jeweils spezifischen Erkrankung.

Eine Vielzahl wissenschaftlicher Institutionen ist im Dialog, um wissenschaftlich basierte Kriterien für die Entwicklung von einer MCID für die verschiedenen Erhebungsinstrumente und damit für die zugrunde liegende Erkrankung zu entwickeln (SISAQOL, IMI u. v. m.).

105 IQWiG. Allgemeine Methoden Entwurf für Version 6.0 vom 05.12.2019

106 vfa. Stellungnahme zum Entwurf der Allgemeinen Methoden Version 6.0 (2020)

107 IQWiG. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden Version 5.0 vom 10.07.2017 2017 [Available from: https://www.iqwig.de/methoden/allgemeine-methoden_version-5-0.pdf?rev=180459, accessed 23-07-2021]

108 Coon CD, Cook, K. F. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res. 2018;27:33-40.

Mit der Implementierung des IQWiG-Methodenpapiers 6.0 (109) hat das IQWiG die Forderung, dass lediglich als validiert geltende oder etablierte MCID als Responsekriterium für Responderanalysen zu verwenden sind (107), aufgehoben und definiert als allgemeingültigen Schwellenwert mindestens 15 % der theoretisch möglichen Skalenspannweite des verwendeten Erhebungsinstruments ohne wissenschaftliche Rationale. Leider bleiben damit auch die bisherigen Ausarbeitungen der genannten Arbeitsgruppen in dem Entwurf des Methodenpapiers unberücksichtigt.

Der G-BA folgt mit der Aufnahme der oben zitierten Passage in die Dossiervorlage dem IQWiG-Methodenpapier 6.0 und lässt die wissenschaftlichen Diskussionen damit ebenfalls unberücksichtigt. Auf diese Weise hält in die Nutzenbewertung eine nicht wissenschaftlich begründete Responseschwelle für die patientenberichteten Endpunkte in zur Zulassung durchgeführten klinischen Studien Einzug.

Darüber hinaus konterkariert der allgemeingültige Schwellenwert von 15 % der Skalenspannweite den bisherigen Gedanken, für verschiedene Therapiegebiete jeweils einen spezifischen Schwellenwert zu validieren, um den verschiedenen Indikationen gerecht zu werden ebenso wie dem Konzept einer MCID, die einen von Patienten empfundenen minimalen klinisch bedeutsamen Unterschied widerspiegelt.

Die Janssen-Cilag GmbH sieht aus den zuvor genannten Gründen den Einzug eines allgemeingültigen Schwellenwertes kritisch.

Durch die Einführung eines für alle Erhebungsinstrumente gleichen Schwellenwertes von 15% besteht das Risiko, dass dieser Schwellenwert für manche Krankheitsentitäten zu hoch angesetzt ist und bei der Durchführung einer Responderanalyse Patienten, die eine spürbare Veränderung wahrnehmen, nicht als Responder detektiert werden, obwohl sie Responder sind. Ein relevanter Patientennutzen bliebe somit unerkannt. Für andere Entitäten kann umgekehrt der Schwellenwert zu niedrig angesetzt sein, sodass Patienten als Responder angesehen werden, obwohl noch keine spürbare Veränderung eingetreten ist.

Bayer Vital GmbH

Die vorgeschlagene „präspezifizierte“ Mindestresponderschwelle von 15% der Skalenbreite entbehrt jeglicher methodischen Fundierung und erscheint als eine unreflektierte Übernahme einer Setzung mit normativem Charakter des IQWiG. Selbst unter der Prämisse, dass es derzeit keine akzeptierten Standards gibt, mit denen die Qualität von Studien bewertet und die Aussagekraft der ermittelten MIDs abgeschätzt werden könne, was auch wiederum zu hinterfragen wäre, würde dieser Mangel durch einen anderen Mangel ersetzt werden. Da dieses Vorgehen weder die spezifischen Skaleneigenschaften noch die Änderungssensitivität der verwendeten Instrumente in seine Betrachtung einfließen lässt, setzt es ein willkürliches Kriterium an, das sich auf Basis einiger weniger herangezogenen Übersichtsarbeiten zu MIDs in spezifischen Krankheitsbildern angeblich (als arithmetischer Mittelwert) ableiten lässt. Letztere beziehen sich allerdings teilweise auf ausgewählte und sehr spezifische Symptome wie bspw. „fatigue“ und „symptoms and functional limitations in individuals with rotator cuff disorders“, sodass deren Ergebnisse nicht allgemeine Geltung über alle Instrumente hinweg genießen können. Letztlich handelt es sich bei diesem Vorgehen des IQWiG erwiesenermaßen um einen Alleingang, der die jahrelange Arbeit von Arbeitsgruppen an Validierungsstudien schlicht negiert.

Ferner stellt sich eher die Frage nach verteilungsbasierten MIDs (bspw. 1/2 SD), falls ankerbasierte MIDs im Rahmen ihrer Validierungsstudien gravierende Mängel aufweisen sollten, um diese Mängel zu heilen. Jedenfalls ist der Ansatz einer exogen festgelegten 15%

109 IQWiG. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden Version 6.0 vom 05.11.2020 2020 [Available from: https://www.iqwig.de/methoden/allgemeine-methoden_version-6-0.pdf?rev=144030, accessed 23-07-2021].

Schwelle der entsprechenden Skalenbreite nicht zielführend, da damit bereits validierte MIDs zu etablierten Instrumenten, die auch für entsprechende Studienplanungen herangezogen wurden, insofern diese niedriger ausfallen, hinfällig werden. […]

Eine Mindestresponderschwelle von 15% der Skalenbreite der entsprechenden Erhebungsinstrumente ist bis dato weder ein internationaler Standard noch scheint es diesbezüglich in den hierfür international aktiven Gremien und Organisationen (Cochrane, HTAi, ISPOR usw.) aus den zur Verfügung stehenden Quellen eine Vorarbeit diesbezüglich gegeben zu haben. Das Kriterium ist somit ein IQWiG Eigenkonstrukt und entspricht in keiner Weise internationalen Standards der Evidenz-basierten Medizin, es sei denn das IQWiG gilt als neuer Maßstab hierfür.

Rein erkenntnistheoretisch ist die Setzung einer Mindestresponderschwelle von 15% der Skalenbreite eingesetzter Erhebungsinstrumente auch deswegen zu hinterfragen, weil sie der (Bio-)Statistik eine normative Qualität zuspricht, die sie epistemiologisch als etablierte Hilfswissenschaft genuin nicht besitzt. Während ankerbasierte MIDs psychometrischen Konstrukten folgen, die am Erkenntnisobjekt (betroffene Patienten) selbst umgesetzt werden, ist die Setzung der 15% Schwelle an der Skalenbreite weder klinisch ableitbar noch psychologisch fundiert. […] Eine idiosynkratische Perzeption der Belastbarkeit von MID Validierungsstudien des Auftragsinstituts des G-BA und daraus folgende Setzungen, die weder international in der EbM-Welt diskutiert, geschweige denn akzeptiert wurden, sollte jedoch keine Grundlage für die Einbringung solcher Kriterien in die Dossiermodulvorlagen des G-BA zur Nutzenbewertung von Arzneimitteln bilden. Vielmehr sollte die Problematik als Anreiz für einen internationalen Dialog fungieren und die entsprechenden etablierten Forschergruppen und Konsortien dazu animieren, neue Validierungsstudien für MIDs durchzuführen, an welchen sich im Nachgang die pharmazeutische Industrie auch orientieren kann. An solch einen Dialog könnte auch die pharmazeutische Industrie mit ihrer Fachexpertise aktiv partizipieren, da sie auch international an mehreren HTA Prozessen involviert ist. Denn die pharmazeutische Industrie wird nicht immer ungerechtfertigte nationale methodische Setzungen mit ihren international ausgerichteten Studien bedienen können. Bis dieses Vorgehen erfolgreich abgeschlossen ist, sollten bereits im Rahmen von (frühen) Nutzenbewertungen akzeptierte MIDs weiterhin akzeptiert werden uns sukzessive bei Vorliegen neuer Validierungsstudien ersetzt werden.

Boehringer Ingelheim Pharma GmbH & Co. KG

1. Bedeutung der wissenschaftlichen und klinischen Gemeinschaft für die Entwicklung von MIDs

Aus Sicht von Boehringer Ingelheim ist die Beteiligung der auf diesem Gebiet tätigen bzw. forschenden WissenschaftlerInnen ein wichtiger Baustein bei der Entwicklung von indikationsspezifischen Responsekriterien bzw. MIDs („Minimal Important Differences“) zur Beurteilung von Patienten-berichteten Endpunkten.

Deshalb möchte sich Boehringer Ingelheim dafür aussprechen, auch weiterhin die Ergebnisse von validierten und international akzeptierten Responsekriterien in der frühen Nutzenbewertung zu berücksichtigen.

Des Weiteren unterstützt Boehringer Ingelheim den Vorschlag, einen akzeptierten Katalog von Bewertungskriterien zu entwickeln, der eine angemessene Beurteilung der Zuverlässigkeit von MIDs erlaubt. Dieser sollte auf Grundlage der bisherigen Empfehlungen und im weiteren gemeinsamen Dialog aus Wissenschaft, Institutionen und Industrie entwickelt werden.

Ecker + Ecker GmbH

Die Ecker + Ecker GmbH sieht [die geplanten Änderungen] kritisch. Schon der Grundgedanke, ein klinisches Relevanzkriterium aus der Skalenspannweite eines Instrumentes abzuleiten, ist nicht sachgerecht. Insbesondere ist jedoch die Herleitung des 15%-Kriteriums nach wie vor

fragwürdig. Dar-über hinaus stellt die mit der allgemeinen Implementierung des 15%-Kriteriums verbundene Ablehnung etablierter und für die Messung spürbarer Änderungen entwickelter MCID eine Abwertung der Patientenperspektive in der Interpretation patientenberichteter Endpunkte dar. Schließlich bleibt die Praktikabilität eines pauschalen, auf alle Skalen und Instrumente angewandten Response-Kriteriums, fraglich.

1 Die Herleitung des 15%-Kriteriums bleibt fragwürdig

Auch nach der erfolgten Würdigung der Stellungnahmen zur Änderung des Methodenpapiers des IQWiG bleibt festzustellen, dass die Herleitung des 15%-Kriteriums nicht evidenzbasiert erfolgte und auf einer zweifelhaften Grundlage beruht. Nach der initialen Kritik durch viele Stellungnehmer hat das IQWiG zwar die erfolgte Literatursuche zu bislang publizierten MCID detaillierter dargestellt. Jedoch erscheint das Vorgehen immer noch willkürlich und ist nicht nachvollziehbar.

So werden sehr niedrige MCID mit der Begründung, sie seien „unrealistisch klein“ nicht mit einbezogen, große MCID jedoch schon. Dabei stellt das IQWiG auch für große MCID fest, dass es „die Frage, ob die Werte am oberen Rand des Spektrums verlässlicher sind, […] aufgrund fehlender akzeptierter Standards zur Qualitätsbewertung, wie auch fehlender Berichtsqualität von MID-Studien, nicht beantworten“ kann [110]. Warum daher große MCID nicht ebenfalls aus der Betrachtung ausgeschlossen werden, bleibt offen. Darüber hinaus hat das IQWiG für den Anspruch, dass 15%-Kriterium würde hinreichend sicher unabhängig von Skala und Kontext für den Patienten relevante Veränderung abbilden, keinerlei Evidenz vorgelegt.

Ohnehin beruht die Herleitung des 15%-Kriteriums auf einer widersprüchlichen Argumentation. Denn die Begründung dafür, ein neues Vorgehen zur Bestimmung einer klinischen Relevanzschwelle zu entwickeln, ist die von IQWiG und G-BA konstatierte unklare Qualität bisheriger MCID. So schreibt der G-BA in den Tragenden Gründen zum Beschluss für dieses Stellungnahmeverfahren:

„Demnach zeigen systematische Zusammenstellungen empirisch ermittelter MIDs, dass zu einzelnen Instrumenten häufig eine Vielzahl von MIDs publiziert werden, die innerhalb eines Erhebungsinstruments große Spannweiten haben können. Ursachlich hierfür können unter anderem die in den Studien eingesetzten unterschiedliche Anker, Beobachtungsperioden oder analytische Methoden sein. Gleichzeitig ist eine anhand methodischer Qualitätskriterien begründete Auswahl empirisch ermittelter MIDs für die Nutzenbewertung derzeit nicht zu treffen“ [111].

Wenn die bisherigen MCID in ihrer Qualität jedoch nicht beurteilbar sind, können sie schlechterdings nicht für eine Herleitung eines übergeordneten Kriteriums herangezogen werden. Daher ist die Herleitung des 15%-Kriteriums durch eine mehr oder minder systematische Zusammenschau bisheriger MCID nicht nachvollziehbar. Genauso gut könnte man im Umkehrschluss folgern, dass die zur Ableitung dieses Kriteriums herangezogen MCID valide sein müssen. Dann aber könnten sie auch direkt als klinisches Relevanzkriterium herangezogen werden.

Die Inkonsistenz dieses Vorgehens spiegelt sich auch in der Tatsache wider, dass MCID, die größer als 15% der Skalenspannweite sind, in der frühen Nutzenbewertung herangezogen werden sollen. Die Anerkennung einer MCID hängt dadurch von ihrem Ausmaß ab und nicht von der zu ihrer Herleitung verwendeten Methodik. Es ist jedoch die Qualität der Methodik, die über die Validität einer MCID entscheidet. Dabei unterscheidet sich die Qualität der zur Ermittlung von MCID angewandten Methodik nicht grundsätzlich zwischen MCID, die

110 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Dokumentation und Würdigung der Anhörung zum Entwurf der Allgemeinen Methoden 6.0. 2020.

111 Gemeinsamer Bundesausschuss (G-BA). Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über die Einleitung eines Stellungnahmeverfahrens zur Änderung der Verfahrensordnung: Änderung der Modulvorlage in der Anlage II zum 5. Kapitel. 2021.

unterhalb von 15% der Skalenspannweite liegen und solchen, die oberhalb dieser Grenze liegen. Wenn MCID über dieser Grenze als valide betrachtet werden können, gilt dies für MCID unterhalb dieser Schwelle natürlich nicht minder.

Astellas Pharma GmbH

Es ist nicht nachvollziehbar, warum etablierte und wissenschaftlich validierte MIDs, deren klinische Relevanz bereits belegt sind, im Rahmen der Nutzenbewertung nicht weiterhin verwendet werden sollten. In diesem Zusammenhang wird um eine detaillierte Begründung und Diskussion der vorgeschlagenen Schwelle von Seiten des G-BAs gebeten. […]

Darüber hinaus kann es für die Beurteilung der klinischen Relevanz keine allgemeingültige Responderschwelle geben, da hierbei jegliche krankheits- und populationsspezifischen Einflüsse sowie Skalencharakteristika ignoriert werden. Dasselbe Instrument kann z.B. in verschiedenen Indikationen unterschiedliche MIDs aufweisen. Auch die Effektrichtung (Verbesserung oder Verschlechterung) hat einen Einfluss auf die MID. Eine Differenzierung zwischen relevanter Verbesserung und Verschlechterung findet bei einer universellen Responderschwelle von 15 % nicht statt. Insbesondere MIDs, die auf Basis desselben Erhebungsinstruments abgeleitet worden sind, weisen für Verschlechterungen tendenziell größere Werte auf als jene für Verbesserungen (Nordin et al. 2016). Dies unterstreicht, dass ein pauschales Vorgehen nicht sinnvoll ist. Des Weiteren besteht die Gefahr, dass zugunsten einer angenommenen höheren Ergebnissicherheit patientenrelevante Veränderungen möglicherweise übersehen werden. […]

Der Weg über eine willkürlich festgesetzte pauschale Responderschwelle ist aus Sicht von Astellas aufgrund der genannten Argumente nicht zielführend. Insgesamt lehnt Astellas daher die vorgeschlagene Änderung ab.

Bewertung

Die Studien zur Bestimmung von MIDs entsprechen in der Regel nicht dem aktuellen Stand der wissenschaftlichen Erkenntnis; es wird auf die Bewertung der Einwände zum Saint-George's Respiratory Questionnaire (SGRQ) auf den Seiten 11 und 12 verwiesen.

Dementsprechend gehen Responderanalysen auf Basis eines Responsekriteriums im Sinne einer MID mit wesentlichen Unsicherheiten bei der Interpretation der klinischen Relevanz des beobachteten Effektes einher.

Es zeigte sich ein allgemeiner Konsens, dass Responderanalysen allgemeine Vorteile gegenüber der Analyse stetiger Daten aufweisen.

Um vor diesem Hintergrund weiterhin die Berücksichtigung von Responderanalysen im Rahmen der Nutzenbewertung zu ermöglichen, wurde vom IQWiG ein neues Vorgehen zur Beurteilung klinischer Relevanzschwellen bei komplexen Skalen unterbreitet.

In Bezug auf die Kritik am methodischen Vorgehen zur Ableitung der Responseschwelle von 15 % ist festzuhalten, dass im IQWiG-Methodenpapier ein Wert von 15 % der Spannweite der jeweiligen Skalen als plausibler Schwellenwert für eine hinreichend sicher spürbare Veränderung empirisch gestützt hergeleitet wurde.

Das Vorgehen wurde im Detail in der Anhörung und Würdigung der Stellungnahmen zum Methodenpapier 6.0 des IQWiG und in der mündlichen Anhörung zur Anpassung der Anlage II.6 zum 5. Kapitel der VerfO erörtert. U.a. erfolgte die Herleitung auf der Grundlage einer fokussierten Recherche nach systematischen Übersichtsarbeiten zu MIDs und unter Berücksichtigung der aktuellen Mindestqualitätskriterien zur Methodik der Ermittlung der MID. Es mussten u.a. folgende Kriterien erfüllt werden: longitudinale Studie, ankerbasierte MID, patientenberichteter Anker, Global-Rating-of-Change(GRC)-Anker, der Cut-off für den GRC-Anker sollte bei minimal, small, little oder höchstens moderate liegen, um die Ermittlung einer MID zu gewährleisten. Die extrahierten MIDs wurden im Verhältnis zur Spannweite der

jeweiligen Skala dargestellt (MID in % der Skalenspannweite). Die Werte am unteren Rand der erhobenen MIDs wurden vom IQWIG u.a. aufgrund der Abgrenzung von spürbarer Veränderung und Messunsicherheit als wenig geeignet eingestuft. Hinsichtlich der Frage, ob die Werte am oberen Rand des Spektrums verlässlicher sind, kommt das IQWiG zu dem Schluss, dass sich diese Frage aufgrund fehlender akzeptierter Standards zur Qualitätsbewertung, wie auch fehlender Berichtsqualität von MID-Studien, nicht beantworten lässt. Jedoch zeige die systematische Betrachtung, dass der wesentliche Anteil empirisch ermittelter MIDs unterhalb von 20 % der jeweiligen Skalenspannweite liegt.

Auf Basis dieser aktuell bestverfügbaren Evidenz erfolgte vom IQWIG eine empirisch gestützte Setzung der Responseschwelle von 15 %, welche als plausibel geeignet angesehen wird, hinreichend sicher eine für Patientinnen und Patienten spürbare Veränderung abzubilden.

Die Herleitung der Responseschwelle von 15 % fand folglich unter Berücksichtigung des aktuellen methodischen Diskurses zu dieser Thematik statt, womit die Festlegung der Responseschwelle auf Basis des aktuellen Standes der wissenschaftlichen Erkenntnis vorgenommen wurde.

2.3.3 Gegenüberstellung von Ergebnissen akzeptierter MIDs aus abgeschlossenen Nutzenbewertungen mit einer Responseschwelle von 15 %

Einwand

Verband Forschender Arzneimittelhersteller e. V.

Simulation zu Auswirkungen des 15-%-Schwellenwerts

Der neue vom IQWiG vorgeschlagene Schwellenwert kann für patientenberichtete Auswertungen einen erschwerten methodischen Nachweis von Vor- und Nachteilen von Therapien bedeuten. Falls Verbesserungen oder Verschlechterungen des patientenzentrierten Befindens bestehen, könnten sie weniger häufig entdeckt und berücksichtigt werden. Die praktischen Auswirkungen der Anwendung des vorgeschlagenen generischen Richtmaßes wurden jedoch seitens des IQWiG weder vor der Änderung seines Methodenpapiers noch danach untersucht.

Seitens der pharmazeutischen Unternehmen (Arbeitsgruppe MID) wurde inzwischen eine umfangreiche Simulation erstellt, um die potenziellen Konsequenzen der Anwendung des 15-%-Schwellenwerts aufzuzeigen. Die Quellcode-offene Web-Applikation ist unter der Webadresse https://htaor.shinyapps.io/midapp/ frei zugänglich und erlaubt Datensimulationen basierend auf kontrollierten und frei modifizierbaren Annahmen [112]. Mit der Hilfe von Simulationen können solche Szenarien, in denen sich maßgebliche Unterschiede in den Responsekriterien ergeben, eindeutig identifiziert werden. Insb. kann damit der Fehler 2. Art (Wahrscheinlichkeit keinen Effekt anzuerkennen, obwohl ein Unterschied vorliegt) und damit die Power der Anwendung des 15-%-Schwellenwerts im Vergleich zu etablierten MID-Schwellen untersucht werden. Die Simulationen weisen zugleich enorme Vorteile gegenüber empirischen Fällen auf, die eine eher geringe wissenschaftliche Aussagekraft haben und keine Aussage darüber erlauben, welcher Analyseansatz aussagekräftiger ist, um die Frage nach Vor- bzw. Nachteilen zu beantworten.

Die Ergebnisse der MID-Arbeitsgruppe weisen mittels der Simulationen darauf hin, dass oft nicht berücksichtigte Kriterien (wie etwa die Schiefe der Verteilung) die Power relevant und abhängig von dem Responsekriterium beeinflussen. Die durchgeführten Simulationen zeigen dabei auch, dass in der deutlichen Mehrzahl der untersuchten Szenarien ein Powerverlust durch die Anwendung der 15-%-Schwelle entsteht. Dies bedeutet, dass sowohl die positiven wie auch negative PRO-Effekte eines neuen Arzneimittels damit oft nur mit geringerer Wahrscheinlichkeit als statistisch signifikant zu entdecken wären als mit etablierten MID-Schwellenwerten. Der Powerverlust der 15-%-Schwelle zeigt sich insb. bei niedrigen bis moderaten Responseraten und einer schiefen Baseline-Verteilung [112, 113].

Dem G-BA wird mit der öffentlich zugänglichen Simulation zugleich eine vollumfängliche Möglichkeit zur Verfügung gestellt, alle denkbaren Szenarien der angedachten Anwendung der 15-%-Schwelle je nach Instrument und Effekten im Rahmen einer wissenschaftlich-statistischen Untersuchung seinerseits zu untersuchen. Die Simulation sollte daher vom G-BA als eine transparente und für alle nachvollziehbare Diskussionsgrundlage verwendet werden.

Beispiele für die Auswirkungen des 15-%-Schwellenwerts in der Praxis der Nutzenbewertung

Die Ergebnisse der o. g. Simulation werden, trotz der noch kurzen Anwendungsdauer der neuen IQWiG-Methodik, bereits durch erste empirische Beispiele aus der Nutzenbewertung bestätigt. Im Verfahren zum Wirkstoff Secukinumab zeigte sich für die Erhebung der Lebensqualität mit dem häufig eingesetzten Instrument SF-36 bei der Anwendung der 15-%-

112 Miller R, Böhm D, Böckmann D, Andreas JO, Pfarr E, Knörzer D, Kupas K, Leverkus FW. Simulation: Einfluss der Responseschwelle in Nutzenbewertungsverfahren (2021)

113 MID-Arbeitsgruppe. MID –One Size Fits All? Powerbetrachtung für Responsekriterien von 10 % bzw. 15 % der Skalenspannweite (2021)

Schwelle (≥ 9,6 Punkte beim psychischen Summenscore MCS bzw. ≥ 9,4 Punkte beim physischen Summenscore PCS) kein statistisch signifikanter Unterschied [114]. Mit der etablierten Schwelle von ≥ 5 Punkten konnte hingegen für MCS ein signifikanter und relevanter Vorteil belegt werden. Zugleich hat der G-BA die Schwelle von ≥ 5 Punkten als hinreichende Annäherung an eine geeignete MID für die Summenskalen des SF-36 akzeptiert. In zwei weiteren Verfahren zu den Wirkstoffen Nivolumab und Ipilimumab zeigte sich für die Erhebung des Gesundheitszustands mit dem breit eingesetzten Instrument EQ-5D VAS bei der Anwendung der 15-%-Schwelle kein statistisch signifikanter Unterschied [115, 116]. Bei der Verwendung von etablierten MID-Schwellen von ≥ 7 Punkten bzw. ≥10 Punkten wurde ein statistisch signifikanter Vorteil zugunsten der Therapie abgeleitet und vom G-BA anerkannt.

Die Praxistauglichkeit des IQWiG-Vorschlags ist damit insgesamt deutlich zu hinterfragen, wenn für die mit am häufigsten eingesetzten PRO-Instrumente und etablierten MID-Schwellen in der Nutzenbewertung relevante Effekte der Arzneimittel nicht mehr aufzuzeigen sind.

Boehringer Ingelheim Pharma GmbH & Co. KG

2. Gegenüberstellung Ergebnisse von Analysen mit international akzeptierten MIDs und mit einer Responseschwelle von 15%

[…] Die innerhalb der letzten zwei Jahren eingereichten Dossiers zur frühen Nutzenbewertung von Boehringer Ingelheim beinhalteten bereits Ergebnisse zur Responseschwelle von 15% der Skalenspannweite.

Somit liegen zum Wirkstoff Nintedanib (Neues Anwendungsgebiet: Chronische progredient fibrosierende interstitielle Lungenerkrankungen; 2020-08-15-D-568) und zum Wirkstoff Jardiance (Neues Anwendungsgebiet: chronische Herzinsuffizienz) zu zwei aktuellen Verfahren vergleichende Daten vor.

3. Beurteilung der Güte von Responsekriterien durch Simulationen

Statistische Untersuchungen zur Güte von Tests basieren auf theoretisch abgeleiteten oder simulierten Verteilungen. Die dabei verwendeten Datensimulationen basieren auf kontrollierten Annahmen, können gezielt auch nur in einzelnen Parametern modifiziert werden und erlauben somit hohe Transparenz und Nachvollziehbarkeit.

Die von der Arbeitsgruppe MID der AG Biostatistik des vfa vorgestellte und öffentlich zugängliche Simulation (https://htaor.shinyapps.io/midapp/) ist aus Sicht von Boehringer Ingelheim geeignet, um potenzielle Konsequenzen der Anwendung eines 15%-Schwellenwertes aufzuzeigen. Mittels der Simulation kann die statistische Power (Wahrscheinlichkeit einen Effekt korrekt anzuerkennen) für die Anwendung des 15%-Schwellenwerts im Vergleich zu etablierten MID-Schwellen verglichen werden. Darüber hinaus können Szenarien identifiziert werden, in denen die Anwendung des 15%-Schwellenwertes im Vergleich zu etablierten MID-Schwellen zu deutlichen Unterschieden führt.

114 G-BA. Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII – Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V Secukinumab (Neubewertung aufgrund neuer wissenschaftlicher Erkenntnisse (Psoriasis-Arthritis)) (18. Februar 2021)

115 G-BA. Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM RL): Anlage XII – Anlage XII – Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V Nivolumab (Neues Anwendungsgebiet: Nicht-kleinzelliges Lungenkarzinom, Kombination mit Ipilimumab und platinbasierter Chemotherapie, Erstlinie) (3. Juni 2021)

116 G-BA. Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM RL): Anlage XII – Anlage XII – Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V Ipilimumab (Neues Anwendungsgebiet: Nicht-kleinzelliges Lungenkarzinom, Kombination mit Nivolumab und platinbasierter Chemotherapie, Erstlinie) (3. Juni 2021)

Boehringer Ingelheim spricht sich dafür aus, die Erkenntnisse der Simulationen bei der Bewertung der Änderung Anlage II.6 (Modul 4) der Verfahrensordnung zu berücksichtigen.

Roche Pharma GmbH

Roche hält den Ansatz, Ergebnisse akzeptierter MIDs aus abgeschlossenen Nutzenbewertungen einer Responseschwelle von 15% gegenüber zu stellen, als nicht zielführend. Eine Auswahl an Resultaten abgeschlossener Nutzenbewertungen kann nicht repräsentativ für diese methodische Fragestellung sein. Das Ergebnis einer solchen Gegenüberstellung birgt ein hohes Verzerrungspotential, da mögliche Unterschiede zwischen generischer Responseschwelle und MID durch verschiedene Faktoren beeinflusst werden können. Roche favorisiert in einer solchen Situation daher Simulationen als die validere Methode zur Beantwortung für diese Fragestellung. Simulationen erlauben unter kontrollierten Bedingungen verschiedene Modifikationen (z.B. versch. Verteilungsannahmen) zu berücksichtigen und somit in verschiedenen Szenarien die Konsequenzen unterschiedlicher Responsekriterien ursächlich untersucht und bewertet werden können (s.a. Begründung AG Biostatistik Anhang I).

Eine solche Simulation wurde von der AG Biostatistik (Anhang II) bereitgestellt um den Einfluss einer 10% und 15% Response-schwelle auf die Power zu untersuchen. In dieser Simulation wurden verschiedene Szenarien durchgespielt unter Berücksichtigung der Faktoren wie Behandlungseffekt (“treatment effect”) und Veränderung der Lebensqualität zu Baseline (“change from baseline”). Es zeigte sich, dass je nach Szenario für die 10% Responseschwelle sowohl ein Power-Gewinn als auch ein Power-Verlust verglichen mit der 15% Responseschwelle möglich war. Allerdings lag in den meisten untersuchten Szenarien eine höhere Power für die 10% Responseschwelle gegenüber den 15% vor. Dieser Power-Verlust durch die vorgeschlagene, pauschale 15% Responseschwelle kann nicht im Sinne der Nutzenbewertung sein.

Aus Sicht von Roche sollten diese Simulationen den Grundstein für weitere Diskussionen legen.

AbbVie Deutschland GmbH & Co. KG

Anhand der bisher publizierten einzelnen Beispiele aus diversen Nutzenbewertungen lassen sich keine generellen Schlussfolgerungen zur Vergleichbarkeit etablierter MID und der generischen Responseschwelle ziehen, da spezifische Beispiele per se und immer nur gesonderte, einzelne, populations- und indikationsspezifische Erkenntnisse liefern. Einen allgemeineren und wissenschaftlich nachvollziehbareren Ansatz bietet die von einer Arbeitsgruppe der pharmazeutischen Unternehmer vorgestellte Simulation (https://htaor.shinyapps.io/midapp/), um die verschiedenen Ansätze zu Responseschwelle miteinander zu vergleichen. Hier zeigte sich teilweise ein hoher Powerverlust bei Anwendung der 15%-Schwelle, der umso ausgeprägter ist, je niedriger die Responseraten und je deutlicher die asymmetrische Verteilung der Baseline-Werte ist (117). Es kann also unabhängig von einzelnen Beispielen nicht auf eine Vergleichbarkeit innerhalb der Nutzenbewertung basierend auf einer generellen Responseschwelle und den etablierten Ansätzen per MID ausgegangen werden.

UCB Pharma GmbH:

Für die Untersuchung des Einflusses von Responsekriterien auf die Power eines Behandlungsvergleichs ist, im Vergleich zu der Sichtung von empirischen Ergebnissen, ein Simulationsansatz eindeutig die Methode der Wahl. Mit der von einer Arbeitsgruppe MID

117 Miller R, Böhm D, Böckmann D, Andreas JO, Pfarr E, Knörzer D, et al. Simulation: Einfluss der Responseschwelle in NutzenbewertungsverfahrenErgebnisse der Simulation. 2021.

entwickelten und öffentlich zugänglichen App (https://htaor.shinyapps.io/midapp/ [118]) können verschiedene Datensituationen simuliert und u.a. Powerberechnungen abgeleitet werden. Mittels dieser Simulationen sind Eigenschaften der 15 %-Schwelle im Vergleich zu einer 10 %-Schwelle, die dem Wert einiger etablierter MIDs entspricht, von der Arbeitsgruppe MID untersucht worden. Die Ergebnisse dieser Simulationen weisen darauf hin, dass oft nicht berücksichtigte Kriterien, wie etwa die Schiefe der Verteilung bei Baseline, die Power relevant und abhängig von dem Responsekriterium beeinflussen. In vielen Fällen zeigt sich ein relevanter Powerverlust in einem Bereich von bis zu 20 % zwischen einer 10 %- und 15 %-Schwelle. Das heißt, dass vorhandene Therapieunterschiede bzgl. einer etablierten MID bei der Anwendung der 15 %-Schwelle möglicherweise nicht mehr erkannt werden. Dieses ist umso relevanter, wenn ein möglicher Schaden einer Therapie beurteilt wird.

Bristol-Myers Squibb GmbH & Co. KGaA

BMS hat sich an einer Arbeitsgruppe beteiligt, die die Bedingungen untersucht hat, unter denen sich die Power (auch bekannt als Teststärke; die Wahrscheinlichkeit mit der tatsächlich existierende Therapieunterschiede entdeckt werden) bei Anwendung einer 15% Responseschwelle im Vergleich zu einer 10% Responseschwelle unterscheidet (119, 120). Hierzu wurde eine frei zugängliche, quellcode-offene Web-Applikation entwickelt. Mittels dieser Applikation wurden Simulationen durchgeführt, welche eine höhere Power bei Verwendung einer Responseschwelle von 10% im Vergleich zu 15% in der deutlichen Mehrheit aller untersuchten Szenarien belegen. Dieses Ergebnis legt nahe, dass die Wahrscheinlichkeit, eine Überlegenheit oder eine Unterlegenheit gegenüber der Vergleichstherapie als „signifikant“ nachzuweisen, mit der vorgeschlagenen 15% Responseschwelle in vielen Fällen kleiner ausfällt. Nach Einschätzung der Arbeitsgruppe kann der Einfluss der Responseschwelle nur dann als gering angesehen werden, wenn unabhängig von der gewählten Responseschwelle eine sehr hohe oder sehr niedrige Power vorliegt. Es hat somit den Anschein, dass es sich bei den 15% der Skalenspannweite um eine konservative Schwelle handelt, die einen Zusatznutzen für die betreffenden Endpunkte nur noch bei erheblichen Effekten zulässt. Dies erscheint fragwürdig vor dem Hintergrund des §5 AM-Nutzen V, in welchem auch eine moderate Verbesserung für einen Zusatznutzen klassifiziert wird.

Im Gegensatz zu den vom G-BA geforderten Nachberechnungen für eine versuchte empirische Evaluation, bei der die Auswirkungen auf die Nutzenbewertung jedoch nur auf Basis einer bestimmten Anzahl von Einzelfällen untersucht werden können, erfordert ein wissenschaftlich fundiertes Vorgehen die systematische Untersuchung relevanter Szenarien, beispielsweise anhand einer computerbasierten Simulationsstudie. Hierbei sollten unter anderem weitere Szenarien einbezogen werden, die bislang nicht in der genannten Web-Applikation berücksichtigt wurden. Auch wissenschaftliche Untersuchungen zur Güte von statistischen Testverfahren basieren nicht auf empirischen Daten, sondern entweder auf algebraisch abgeleiteten oder simulierten Verteilungen. Analysen konkreter Datensätze liefern zunächst zwar scheinbar klare Testergebnisse (z.B. anhand der p-Werte), diese lassen sich jedoch nicht verallgemeinern. Letztendlich wird nur indirekt (durch den statistischen Test) berücksichtigt, dass es sich bei den verwendeten Daten um Zufallsstichproben handelt. Werden beispielsweise verschiedene Responsekriterien wie vorgeschlagen auf den gleichen Datensatz angewendet, so lässt sich aufgrund der p-Werte oder anderer statistischer Kenngrößen nicht ableiten, welcher Analyseansatz der validere ist. Datensimulationen dagegen basieren auf klar kontrollierten Annahmen und können gezielt auch nur in einzelnen Parametern modifiziert

118 Andreas JO, Böckmann D, Böhm D, Knoerzer D, Kupas K, Leverkus F, Miller R, Pfarr E (2021): R Shiny App zur Simulation von MID Szenarien. [Zugriff 08.07.2021] URL: https://htaor.shinyapps.io/midapp/.

119 Miller, R., Böhm, D., Böckmann, D., Andreas, J.-O., Pfarr, E., Knörzer, D., Kupas, K., & Leverkus, F.-W. Simulation: Einfluss der Responseschwelle in Nutzenbewertungsverfahren [09.07.2021]

120 Miller, R., Böhm, D., Böckmann, D., Andreas, J.-O., Pfarr, E., Knörzer, D., Kupas, K., & Leverkus, F.-W. MID – One size fits all? Powerbetrachtung für Responsekriterien von 10% bzw. 15% der Skalenspannweite [09.07.2021]

werden, so dass maximale Transparenz und Nachvollziehbarkeit hinsichtlich der resultierenden Aussagen besteht.

Der Umgang mit patientenberichteten Daten birgt – wie in diesem Jahr auch von Experten bei der Veranstaltung „IQWiG im Dialog“ dargestellt – vielfältige Herausforderungen und bislang verfügbare Methoden sind zweifellos mit Schwächen und Limitationen behaftet. Die Berücksichtigung der Patientenperspektive bei der Beurteilung von Lebensqualitätsdaten ist von entscheidender Bedeutung, weshalb eine Verbesserung bzw. Weiterentwicklung der bestehenden Methodik unerlässlich ist. Die normative Festlegung einer pauschalen Responseschwelle auf einen pauschalen Wert, der für alle Lebensqualitätsfragebögen gleichermaßen gelten soll, unabhängig von der abgefragten Domäne und der Komplexität des Fragebogens, liefert jedoch keine Perspektive zur Behebung bestehender methodischer Schwächen im Umgang mit patientenberichteten Endpunkten, sondern würde diesbezüglich vielmehr die wissenschaftliche Weiterentwicklung im Rahmen der Nutzenbewertung unterbinden. Die tatsächliche Relevanz festgestellter Verbesserungen oder Verschlechterungen für Patientinnen und Patienten bleibe durch eine pauschale Responseschwelle weitgehend unberücksichtigt. BMS ist daher der Ansicht, dass ein Kriterienkatalog für die Beurteilung von Responseschwellen unter Einbindung von Wissenschaftlern und Patientenvertretern erarbeitet werden sollte. Zudem sollte die wissenschaftliche Diskussion zu alternativen Auswerteverfahren vorangetrieben werden. Bis es einen wissenschaftlichen Konsens gibt, sollten im Nutzenbewertungsverfahren weiterhin die wissenschaftlich etablierten und in den Studien vordefinierten Responsekriterien Anwendung finden.

IQVIA Commercial GmbH & Co. OHG

IQVIA verfügt nicht über publizierbare Studiendaten, um eine solche Gegenüberstellung vornehmen zu können und hält eine solche Gegenüberstellung aus methodischen Gründen auch nicht für zielführend, um die Auswirkungen einer höheren Responseschwelle zu untersuchen (121). Stattdessen möchten wir auf die Ergebnisse einer Simulationsstudie zur Untersuchung der Auswirkung der höheren Responseschwelle aufmerksam machen. Mit Hilfe der von der „MID Arbeitsgruppe“ entwickelten und öffentlich zugänglichen App (https://htaor.shinyapps.io/midapp/) wurden verschiedene Datensituationen simuliert und u.a. Powerberechnungen durchgeführt. Die Ergebnisse der Untersuchung zeigten, dass bei einer Vielzahl der untersuchten Szenarien, die Verwendung der 15%-Responseschwelle teilweise mit einem hohen Powerverlust einhergeht (122). Dies gilt insbesondere bei Szenarien, in denen der Anteil der Responder gering oder moderat war, sowie bei schiefen, also nicht-symmetrischen Ausgangsverteilungen der Scorewerte. Power-Unterschiede zugunsten der 15%-Responseschwelle fielen in ihrer Stärke relativ geringer aus und wurden in einem deutlich kleineren Teil der untersuchten Szenarien gefunden (123).

GlaxoSmithKline GmbH & Co. KG

Aus Sicht von GSK ist dieses Vorgehen einer selektiven empirischen Nachanalyse nicht zielführend, da die Ergebnisse von einer Vielzahl von Faktoren beeinflusst werden. Dabei wird nur indirekt (durch den statistischen Test) berücksichtigt, dass die verwendeten Daten Zufallsstichproben einer zugrundeliegenden „wahren“ statistischen Verteilung sind. Werden verschiedene Responsekriterien auf den gleichen Datensatz angewendet, so lässt sich aufgrund der p-Werte oder anderer statistischer Kenngrößen nicht ableiten, welcher Analyseansatz der validere ist, um die Frage nach einem Unterschied zu beantworten.

121 MID AG 2021a. Appendix 1 - Begründung des Simulationsansatzes.
122 MID AG 2021b. Appendix 2 - Ergebnisse der Simulation.
123 MID AG 2021c. Appendix 3 - MID - One size fits all. Powerbetrachtungen in Abhängigkeit der Responseraten.

Ein sinnvolleres und in der Wissenschaft etabliertes Vorgehen besteht darin, die Eigenschaften des 15% Schwellenwertes in kontrollierten Simulationsstudien zu untersuchen. Eine Arbeitsgruppe der pharmazeutischen Industrie (AG MID) hat sich mit diesem Ansatz intensiv beschäftigt und die Ergebnisse bereits vorgestellt (124). Des Weiteren stellt die Arbeitsgruppe einen freien Zugang zu der Simulation zur Verfügung (Webadresse: https://htaor.shinyapps.io/midapp/). Die Ergebnisse der Arbeitsgruppe weisen darauf hin, dass oft nicht berücksichtigte Kriterien wie etwa die Schiefe der Verteilung die Power relevant und abhängig von dem Responsekriterium beeinflussen.

Konkret führt die Anwendung der 15%-Schwelle dabei häufig zu einem hohen Powerverlust.

GSK begrüßt dieses Vorgehen ausdrücklich und ist davon überzeugt, dass dieser Ansatz den üblichen wissenschaftlichen Standards genügt.

Pfizer Pharma GmbH

Im Folgenden wird begründet, warum [es] […] seitens Pfizer als nicht zielführend angesehen und daher nicht umgesetzt wird [Ergebnisse akzeptierter und neuer Responseschwellen gegenüberzustellen].

Wissenschaftlich-statistische Untersuchungen zur Güte von Tests basieren nicht auf empirischen Daten, sondern auf theoretisch abgeleiteten oder simulierten Verteilungen. Die Qualität von Simulationsergebnissen ist dabei im Kontext der Konsistenz bei variierenden Annahmen sowie entsprechender theoretischer Ableitungen bzw. Begründungen einzuordnen. Datensimulationen Seite basieren auf kontrollierten und klaren Annahmen, können gezielt auch nur in einzelnen Parametern modifiziert werden und erlauben somit maximale Transparenz und Nachvollziehbarkeit.

Analysen konkreter Datensätze liefern zunächst zwar scheinbar klare Testergebnisse (z.B. anhand der p-Werte), werden aber durch eine Vielzahl von größtenteils unbekannten Faktoren beeinflusst. Letztendlich wird nur indirekt (durch den statistischen Test) berücksichtigt, dass die verwendeten Daten Zufallsstichproben einer zugrundeliegenden „wahren“ Verteilung sind. Werden beispielsweise verschiedene Responsekriterien wie vorgeschlagen auf den gleichen Datensatz angewendet, so lässt sich aufgrund der p-Werte oder anderer statistischer Kenngrößen nicht ableiten, welcher Analyseansatz der validere ist, um die Frage nach einem Unterschied zu beantworten.

Mit einem statistischen Test kann sowohl der Fehler erster Art (Wahrscheinlichkeit einen Effekt zu postulieren, obwohl kein Unterschied vorliegt) als auch der Fehler zweiter Art (Wahrscheinlichkeit keinen Effekt zu postulieren, obwohl ein Unterschied vorliegt) kontrolliert werden. Letzteres wird auch unter dem Begriff „Power“ beschrieben, also die Wahrscheinlichkeit einen wahren Unterschied als „signifikant“ erkennen zu können. Wendet man nun zwei unterschiedliche Responsekriterien auf den gleichen Datensatz an, so bedeutet ein kleinerer p-Wert für eines der beiden Kriterien nicht, dass dieses Kriterium eine validere Entscheidung bezüglich der zu testenden Hypothese liefert. Simulationen erlauben dagegen, unter Kontrolle des angenommenen Effektes die Power abzuleiten. Dabei können verschiedene Modifikationen (z.B. schiefe statt symmetrischer Basis-Verteilungen) berücksichtigt und untersucht werden. Somit können die Szenarien, in denen sich maßgebliche Unterschiede in den Responsekriterien ergeben, klar identifiziert werden.

Zusammenfassend ergibt sich daher für die Fragestellung des Einflusses von Responsekriterien ein eindeutiger Vorteil von Simulationen im Vergleich zu „empirischen“ Ergebnissen, die keinen oder nur einen sehr geringen wissenschaftlich begründbaren Beitrag zur Fragestellung liefern können. Mittels Simulationen können systematisch Eigenschaften

124 Andreas J-O. Patientenberichtete Endpunkte: Wie können aussagekräftige minimale patientenrelevante Unterschiede hergeleitet werden? 2021 20.07.2021. Available from: https://www.iqwig.de/veranstaltungen/2021_iqwig_im_dialog_jens_otto_andreas.pdf?rev=209373.

solcher Situationen identifiziert werden, in denen eine Responseschwelle von 15% anderen Responseschwellen unter- oder überlegen ist.

Seitens der pharmazeutischen Unternehmen (Arbeitsgruppe MID) wurde inzwischen eine umfangreiche Simulation erstellt, um die potenziellen Konsequenzen der Anwendung des 15%-Schwellenwerts aufzuzeigen. Die Quellcode-offene Web-Applikation ist unter der Webadresse https://htaor.shinyapps.io/midapp/ frei zugänglich (125) und ermöglicht es solche Szenarien, in denen sich maßgebliche Unterschiede in den Responsekriterien ergeben, eindeutig zu identifizieren. Insbesondere kann damit die Power der Anwendung des 15-%-Schwellenwerts im Vergleich zu etablierten MID-Schwellen unter verschiedenen Eigenschaften von Studiendesign und Zielpopulation untersucht werden. Die Ergebnisse der MID-Arbeitsgruppe zeigen mittels der Simulationen, dass oft nicht berücksichtigte Eigenschaften (wie etwa die Schiefe der Verteilung) die Power relevant und abhängig von dem Responsekriterium beeinflussen. Die durchgeführten Simulationen zeigen dabei auch, dass in der deutlichen Mehrzahl der untersuchten Szenarien ein Powerverlust durch die Anwendung der 15-%-Schwelle entsteht. Dies bedeutet, dass sowohl die positiven wie auch negative PRO-Effekte eines neuen Arzneimittels damit oft nur mit geringerer Wahrscheinlichkeit als statistisch signifikant zu entdecken wären als mit etablierten MID-Schwellenwerten. Der Powerverlust der 15-%-Schwelle zeigt sich insbesondere bei niedrigen bis moderaten Responseraten und einer schiefen Baseline-Verteilung (125, 126). Die Heat Map in Abbildung 1 zeigt die Ergebnisse der Power-Unterschiede einer 10%- zu einer 15%- Responseschwelle in Abhängigkeit der Veränderung zur Baseline und des Behandlungseffekts. Die Abbildung verdeutlicht, dass die Wahrscheinlichkeit mit der eine Power-Überlegenheit bei Verwendung einer 10% bzw. 15% Responseschwelle zu erwarten ist, maßgeblich davon abhängt, welcher Teil des untersuchten Parameterraums in der Praxis abgedeckt wird. Dieser Frage sollte vor der standardmäßigen Verwendung des generischen 15%-Schwellenwertes im Rahmen eines fortgesetzten wissenschaftlichen Diskurses nachgegangen werden. […]

Dem G-BA wird mit der App eine vollumfängliche Möglichkeit zur Verfügung gestellt, alle denkbaren Szenarien der angedachten Anwendung der 15%-Schwelle je nach Instrument und Effekten im Rahmen einer wissenschaftlich-statistischen Analyse seinerseits zu untersuchen. Dies sollte als eine transparente und für alle nachvollziehbare Diskussionsgrundlage verwendet werden. Nachberechnungen „empirischer“ Datensätze sind dagegen nicht geeignet die wissenschaftliche Fragestellung adäquat und vollständig zu adressieren.

Bewertung

Es ist davon auszugehen, dass der Nachweis eines Effektes nicht von einer spezifischen MID abhängt (deren Validierung zudem aktuell in der Regel nicht dem aktuellen Stand der wissenschaftlichen Erkenntnis entspricht), sondern, dass ein Effekt auch mit dem neu vorgeschlagenen Wert von 15 % der Spannweite der jeweiligen Skalen nachweisbar ist.

Im Zusammenhang mit dem vom IQWiG unterbreiteten neuen Vorgehen zur Beurteilung klinischer Relevanzschwellen bei komplexen Skalen wurde in den Stellungnahmen zum Entwurf der Allgemeinen Methoden 6.0 – wie auch im vorliegenden Stellungnahmeverfahren – eingebracht, dass der neu vorgeschlagene Wert von 15 % der Spannweite der jeweiligen Skalen den Nachweis von Vor- und Nachteilen von Therapien in patientenberichteten Endpunkten erschweren kann.

Um diesen Themenkomplex im Rahmen des vorliegenden Stellungnahmeverfahrens aufzugreifen, wurden die Stellungnehmenden, die über Studiendaten verfügen, bei denen Responderanalysen im Sinne einer MID vom G-BA in abgeschlossenen Nutzenbewertungen

125 MID Arbeitsgruppe. MID - One SIze Fits All? Powerbetrachtung für Responskriterien von 10% bzw. 15% der Skalenspannweite 2021.

126 Miller R, Böhm D, Böckmann D, Andreas J, Pfarr E, Knörzer D, et al. Simulation: Einfluss der Responseschwelle in Nutzenbewertungsverfahren 2021.

berücksichtigt wurden, gebeten, eine Gegenüberstellung dieser Ergebnisse mit denen einer Responseschwelle von 15 % der Skalenspannweite des Instruments in das Stellungnahmeverfahren einzubringen.

Von den Stellungnehmenden wurde keine Gegenüberstellung von Ergebnissen akzeptierter MIDs aus Beschlüssen des G-BA mit einer Responseschwelle von 15 % vorgenommen. Begründet wurde dies u.a. damit, dass ein derartiges Vorgehen aus methodischen Gründen nicht zielführend ist. So sei z.B. eine Auswahl an Resultaten abgeschlossener Nutzenbewertungen nicht repräsentativ für die zugrundeliegende methodische Fragestellung und bürge ein hohes Verzerrungspotential, da mögliche Unterschiede zwischen der Responseschwelle von 15 % und MID durch eine Vielzahl von Faktoren beeinflusst werden können.

Stattdessen wurde von den Stellungnehmenden auf Simulationen abgestellt. Simulationen erlaubten laut den Stellungnehmenden demnach u.a., unter kontrollierten Bedingungen verschiedene Modifikationen zu berücksichtigen und somit in verschiedenen Szenarien die Konsequenzen unterschiedlicher Responsekriterien ursächlich zu untersuchen und zu bewerten.

Die im Stellungnahmeverfahren diskutierte Simulation wurde von einer Arbeitsgruppe der pharmazeutischen Unternehmer bereitgestellt (https://htaor.shinyapps.io/midapp/), um den Einfluss einer 10 % und 15 % Responseschwelle auf die Power zu untersuchen.

In der Simulation zeigte sich laut den Stellungnehmenden, dass je nach Szenario für die 10 % Responseschwelle sowohl ein Power-Gewinn als auch ein Power-Verlust verglichen mit der 15 % Responseschwelle möglich war. Jedoch lag in den meisten untersuchten Szenarien eine höhere Power für die 10 % Responseschwelle gegenüber der 15 % Responseschwelle vor. Dies galt insbesondere bei Szenarien, in denen der Anteil der Responder gering oder moderat war, sowie bei schiefer Baseline-Verteilung. Power-Unterschiede zugunsten der 15 %-Responseschwelle fielen in ihrer Stärke relativ geringer aus und wurden in einem deutlich kleineren Teil der untersuchten Szenarien gefunden.

Im Rahmen der Bewertung der von den Stellungnehmenden vorgelegten Simulations-Untersuchengen ist jedoch festzustellen, dass unklar bleibt, ob die Powerverschiebungen, die sich in der Simulationsstudie zeigen, Auswirkungen auf die Nutzenbewertung haben. So bleibt offen, in welchen Arealen des Parameterraums man sich in der Praxis – im Rahmen der Nutzenbewertung – bewegt, z.B. in wie vielen Dossiers und patientenrelevanten Endpunkten schiefe Baseline-Verteilungen vorliegen. Zur Beantwortung der Frage, welche in der Simulationsstudie identifizierten Konstellationen praxisrelevant sind, ist eine entsprechende Empirie notwendig. Zudem stehen den Simulationsszenarien mit einem Powerverlust aufgrund der 15 % Responseschwelle Szenarien gegenüber, in denen es zu einem Powergewinn kommt.

Abgesehen von den hier beschriebenen methodischen Unsicherheiten ist nicht nachvollziehbar, dass eine Simulation aussagekräftiger ist als ein Ansatz über konkrete Beispiele. Es wurden keine konkreten Beispiele vorlegt, die das nun durch IQWiG und G-BA vorgeschlagene Vorgehen infrage stellen. Auf der anderen Seite zeigen die Erfahrungen in den Nutzenbewertungsverfahren seit Anpassung des Methodenpapiers, dass es sich um ein praktikables Vorgehen handelt. In diesem Zusammenhang wird auch auf die Ausführungen zum Aufbau der EORTC-Fragebögen und der sich daraus ergebenden fehlenden praktischen Konsequenzen eines Wechsels des Responsekriteriums von 10 Punkten auf 15 % für diese Fragebögen auf den Seiten 17 und 18 verwiesen. Die von den Stellungnehmenden vorgelegte Simulation berücksichtigt den tatsächlichen Aufbau der in der Wissenschaft etablierten EORTC-Fragebögen nicht und ist daher für Aussagen, welche Auswirkungen unterschiedliche Responsekriterien auf die Ergebnisse zu diesen Fragebögen haben, ungeeignet.

Insgesamt kann aus der Simulationsstudie nicht abgeleitet werden, dass der neu vorgeschlagene Wert von 15 % der Spannweite der jeweiligen Skalen den Nachweis von Vor- und Nachteilen von Therapien in patientenberichteten Endpunkten in relevantem Ausmaß erschwert. Die Ergebnisse der Simulationsstudie stehen der Anpassung der Anlage II.6 zum 5. Kapitel der VerfO somit nicht entgegen.

2.3.4 Weitere Anmerkungen

Einwand

Implementierung der Responseschwelle von 15 % in die Modulvorlage

IQVIA Commercial GmbH & Co. OHG

Verlust an Flexibilität und Behinderung weiterer Forschung zu PRO-Instrumenten

Unabhängig von den allgemeinen Kritikpunkten an der Methodik erscheint die Implementierung der 15%-Responseschwelle in der Modulvorlage ungewohnt starr und ermöglicht keine flexible, an Skalen oder Indikationen angepasste Handhabung. In der Modulvorlage wurden Spezifikation für Darstellungsformen und Auswertungsarten für Endpunkte bislang beispielhaft genannt (z.B. Ereigniszeitanalysen für unerwünschte Ereignisse bei unterschiedlichen Beobachtungsdauern). Abweichend davon gibt der G-BA für die Ergebnisdarstellung zu PRO jetzt exakt zwei Möglichkeiten vor

1. die stetige Darstellung mit Relevanzbewertung anhand von Hedges'g (bei einem vorgegebenen Irrelevanzbereich von 0,2)

2. Responderanalysen mit einem Responsekriterium von exakt 15% der Skalenspannweite

Erstgenannte Darstellungsart beinhaltet mit der Relevanzbewertung durch Hedges' g eine Methodik, die bei ihrer Einführung im Methodenpapier 4.0 nicht unkritisch gesehen wurde und generell nicht geeignet ist, einen Zusatznutzen auf Endpunktebene in seinem Ausmaß einzustufen (zu quantifizieren). Mit der aktuellen Änderung fände diese Form der Relevanzbewertung erstmals Eingang in die Verfahrensordnung. IQVIA sieht einen größeren Vorteil in einer holistischen Betrachtung der Ergebnisse zu PRO, die sowohl den longitudinalen Charakter der Daten berücksichtigt, als auch die Möglichkeit von Sensitivitätsanalysen einräumt, sowie eine differenzierte Betrachtung von Skalenspezifika wie etwa der Berücksichtigung des Wertebereichs möglicher Änderungen.

Zeitgleich mit der Veröffentlichung des IQWiG Methodenpapiers veröffentlichte eine Gruppe von Wissenschaftlern rund um Gordon Guyatt ein Instrument zur Bewertung der Zuverlässigkeit der Schätzung von ankerbasierten MIDs (127). Die vorgeschlagenen Kriterien sind teilweise im Einklang mit dem wissenschaftlichen Konsens (ankerbasierte Verfahren, longitudinale Daten) – andere haben bereits zu weiteren Diskussionen und Kommentaren in der wissenschaftlichen Community angeregt (128)). Eine Fortsetzung des wissenschaftlichen Austauschs wäre wünschenswert und sollte nicht durch die Festlegung eines singulären "one-size-fits-all"-Kriteriums ausgebremst werden.

Bewertung

Vor dem Hintergrund der Änderungen der methodischen Anforderungen an die Dossiererstellung in Verbindung mit dem bisherigen Vorgehen und den Erfahrungen des G-BA mit der Nutzenbewertung nach § 35a SGB V sollen mit der Anpassung der Modulvorlage Unsicherheiten der pharmazeutischen Unternehmer in der Dossiererstellung im Hinblick auf die Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen vermieden werden.

127 Devji T., Carrasco-Labra A., Qasim A. et al. 2020a. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ (Clinical research ed.) 369, S. m1714.

128 Devji T., Carrasco-Labra A., Qasim A. et al. 2020b. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study: All rapid responses. Verfügbar unter: https://www.bmj.com/content/369/bmj.m1714/rapid-responses, abgerufen am: 14.07.2021.

Unbenommen der aktuellen und anhaltenden wissenschaftlichen Diskussionen um Kriterien für die Entwicklung von einer MID und möglicher zukünftiger Standards ermöglicht die Anwendung eines Responsekriteriums von mindestens 15 % der Skalenspannweite zum jetzigen Zeitpunkt die Berücksichtigung von Responderanalysen im Rahmen der Nutzenbewertung.

Einwand

IQVIA Commercial GmbH & Co. OHG

Konsequenz einer höheren Responseschwelle auf die Bewertung des Ausmaßes des Zusatznutzens

Es bleibt unklar, inwieweit das IQWiG seine Bewertung zum Ausmaß des Zusatznutzens (Quantifizierung des Zusatznutzens) angesichts einer höheren Responseschwelle anpasst. Nach Einschätzung des IQWiG ist die 15%-Schwelle in der Lage, nicht mehr nur eine „minimale spürbare“, sondern eine „hinreichend sicher spürbare“ Veränderung abzubilden. Folglich dürften alle Veränderungen oberhalb der 15%-Schwelle eine bedeutsame Veränderung markieren d.h. eine Veränderung, die über eine "irrelevante" Veränderung hinausgeht. Eine Statistische Signifikanz in den Responder-Auswertungen basierend auf der 15%-Schwelle müsste – auch bei nicht-schweren Symptomen - mindestens als „geringer“ Zusatznutzen gewertet werden. Das ist derzeit, zumindest gemäß IQWiG-Methodik, so nicht vorgesehen. Eine Einordnung der Konsequenzen einer höheren Responseschwelle im Zusammenhang mit „irrelevanten Effekten“ seitens des G-BA wäre daher wünschenswert.

Bewertung

Dies ist Teil der Bewertungsentscheidung des G-BA.

Einwand

GlaxoSmithKline GmbH & Co. KG

Klarstellung zu „komplexe Skalen“

In der vom G-BA vorgeschlagenen Änderung der Dokumentvorlage zu Modul 4 heißt es im Abschnitt 4.3.1.3.1: „Die Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen soll nach folgendem Vorgehen erfolgen:“ Es geht aus der Formulierung „komplexe Skalen“ jedoch nicht eindeutig hervor, auf welche Skalen sich diese Kriterien konkret beziehen. Hier bedarf es aus unserer Sicht noch einer Klarstellung.

Astellas Pharma GmbH

[…] Eine deutliche Limitation der Responderschwelle von 15 % besteht bei Skalen mit unbegrenzter Spannweite. Für diese lässt sich keine Responderschwelle ableiten und somit ist in diesen Fällen die vorgeschlagene Methode gänzlich ungeeignet. […]

Bewertung

Die Anpassung der Anlage II.6 zum 5. Kapitel der VerfO adressiert komplexe Skalen (insbesondere psychometrische Skalen), deren Ergebnisse nicht ohne Weiteres interpretiert werden können. Das Responsekriteriums von 15 % der Skalenspannweite soll nicht bei Skalen eingesetzt werden, deren Ergebnisse in natürlichen Einheiten ausgedrückt werden und die deshalb einer inhaltlichen Bewertung der Relevanz eines Effekts zugänglich sind.

Bei Skalen mit unbegrenzter Spannweite kann das Responsekriterium von 15 % der Skalenspannweite nicht herangezogen werden. Da komplexe Skalen stets über einen Maximalwert verfügen, kann das von den Stellungnehmenden angeführte Problem der Limitation der Responderschwelle von 15 % bei Skalen mit unbegrenzter Spannweite nicht auftreten.

2.3.5 Vorgeschlagene Änderungen

Amgen GmbH

Amgen ist der Auffassung, dass die vom G-BA festgestellte und vom IQWiG angwandte fortwährende Akzeptanz der klinisch relevanten und etablierten MID von ≥ 10 Punkten beim EORTC QLQ-C30 und seinen validierten, krankheitsspezifischen Ergänzungsmodulen in die Modul 4-Formatvorlage unter 2. aufgenommen werden sollte: „ […] Bei der Auswertung des EORTC QLQ-C30 Fragebogens und dessen validierten, krankheitsspezifischen Ergänzungsmodulen wird weiterhin die etablierte und validierte MID von ≥ 10 Punkten für die frühe Nutzenbewertung herangezogen."

Weiterhin sollten auch über den EORTC QLQ-C30 und dessen Ergänzungsmodule hinaus, validierte bzw. etablierte und bereits in früheren Verfahren akzeptierte Responderkriterien validierter PRO-Instrumente Gültigkeit behalten. Das vom IQWiG geforderte pauschale Responderkriterium erscheint wissenschaftlich nicht hinreichend begründet und lässt generell die Besonderheit unterschiedlicher Therapiegebiete und Skalencharakteristika unbeachtet.

Roche Pharma GmbH

A) Aus Sicht von Roche sollten alle bisher vom GBA akzeptierten MIDs weiterhin als valide erachtet werden und für die Nutzenbewertung herangezogen werden. Diese müssen Vorrang vor jeglicher pauschalen Responseschwelle haben.

B) Eine Responseschwelle sollte nur in Situationen, in denen (noch) keine akzeptierte MID vorliegt, als temporäre Lösungen betrachtet werden. Allerdings dürfen die Responseschwellen in solchen Situationen nicht willkürlich und generisch ausgewählt werden, sondern auf Basis von Simulationsverfahren.

C) Roche hält die Verwendung von Ergebnissen aus abgeschlossenen Nutzenbewertungen für die Bewertung der Responseschwelle von 15% für nicht angebracht. Stattdessen sollte ein Simulationsbasiertes Verfahren verwendet werden. Ein Beispiel solcher Simulation wurde bereits von der AG Biostatistik durch-geführt (Anhang I). Aus Sicht von Roche sollte dies die Grundlage für weitere Diskussionen sein.

D) Roche hält zur Lösung von diesen und ähnlich gelagerten Fragestellungen ein Gremium für unabdingbar, dass aus IQWiG, GBA, Akademie, und Industrievertretern besteht – mit dem Ziel, einen internationalen und wissenschaftlich fundierten Kriterien-katalog zu etablieren, der für MIDs im Rahmen von Nutzenbewertungen regelhaft anzuwenden ist. Bis zur Fertigstellung dieses Kriterienkatalogs sollten alle etablierte und akzeptierten MIDs weiterhin vom GBA für die Nutzenbewertung herangezogen werden.

GlaxoSmithKline GmbH & Co. KG

Das vorgeschlagene Vorgehen, ein pauschales Responsekriterium von 15 % der Skalenspannweite für alle Erhebungsinstrumente und für alle Indikationen anzuwenden, ist aus Sicht von GSK wissenschaftlich nicht valide begründet. Die fehlende wissenschaftliche Validität sowie die Nicht-Berücksichtigung von Patienten-Perspektiven wird auch von den Entwicklern zweier etablierter Instrumente (SGRQ und SF-36 bzw. SF-12) bestätigt. GSK lehnt daher die vorgeschlagene Änderung der Modulvorlage in Abschnitt 4.3.1.3.1 ab und schlägt stattdessen folgendes Vorgehen vor:

Anstatt sich auf ein pauschales, einfach anzuwendendes, aber wissenschaftlich höchst fragwürdiges Kriterium zu fixieren, sollte mit Wissenschaftlern und Patientenvertretern ein Kriterienkatalog für Standards bei der Etablierung von MIDs aufgestellt werden. In einer breit angelegten Diskussion sollten auch die Eigenschaften von anderen Analyseverfahren (z.B. Fläche unter der Kurve) zur Validierung von MIDs untersucht werden. Dadurch würde eine inhaltliche Weiterentwicklung zur adäquaten Auswertung von patientenrelevanten

Endpunkten, unter Berücksichtigung von Endpunkt-spezifischen patientenrelevanten Kriterien, im Sinne der evidenzbasierten Medizin erfolgen. Bis zum Abschluss dieser Diskussion sollten die etablierten MIDs, die bereits in Nutzenbewertungsverfahren vom G-BA akzeptiert wurden, auch weiterhin Berücksichtigung und Akzeptanz finden. Für Endpunkte, bei denen bisher noch kein MID etabliert wurde, kann – im Sinne einer pragmatischen Übergangslösung -vorübergehend das vorgeschlagene 15%-Kriterium herangezogen werden.

MSD Sharp & Dohme GmbH

Auch wenn die Responderschwelle mit einem 15% Schwellenwert die Analyse für die Nutzenbewertung generalisieren und somit stark vereinfachen würde, wird dabei ein wissenschaftlich höchst umstrittenes Responsekriterium gewählt; anstatt eine indikations- und studienspezifische Abwägung der Erhebungsinstrumente und der herangeführten verwendeten Relevanzschwellen durchzuführen. MSD ist sich der Wichtigkeit der inhaltlichen, wie wissenschaftlichen Weiterentwicklung zur PRO Bewertung bewusst, hat jedoch erhebliche Bedenken, dass mit der Verabschiedung eines 15% „one-size-fits-all" Kriteriums eine objektive Diskussion gehemmt wird.

Zur Förderung der wissenschaftlichen PRO Bewertungsweiterentwicklung sollten dementsprechend zukünftige Lösungsvorschläge in Zusammenarbeit mit erfahrenen Wissenschaftler*innen und auch Patientenvertreter*innen entwickelt werden. Zudem sollte man eine breite angelegte Diskussion bezüglich anderer Auswerteverfahren (zeitabhängige Verfahren, Fläche unter der Kurve etc.) bedenken.

Novartis Pharma GmbH

[…] Unter der Voraussetzung, dass (i) gut validierte und etablierte MIDs vorliegen und (ii) diese MIDs in den zur Bewertung herangezogenen Studien präspezifiziert wurden, sollen die entsprechenden MIDs grundsätzlich weiterhin für die Nutzenbewertung anerkannt werden, auch wenn diese nicht mindestens 15 % der Skalenspannweite entsprechen.

Die Festlegung von Mindeststandards an Validierungsstudien ist aktuell Gegenstand der Forschung. Somit ist zu erwarten, dass in Zukunft mehr gut validierte Responderschwellen verfügbar sein werden. Ein mögliches Vorgehen, um hinreichend validierte MIDs in der Nutzenbewertung zu berücksichtigen, wäre die Erfassung solcher PROs und MIDs in einer dynamischen Datenbank.

In Fällen, in denen keine validierten MIDs vorliegen, kann grundsätzlich eine feste Grenze von x % sinnvoll sein. Das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) leitet in seinem Methodenpapier 6.0 aus mehreren Übersichtsarbeiten zu MIDs eine Schwelle von 15 % her. Die in bisherigen Verfahren nach dem Arzneimittelmarktneuordnungsgesetz (AMNOG) anerkannten MIDs sind jedoch mehrheitlich deutlich kleiner als 15 % der jeweiligen Skalenspannweite (z. B. bei EQ-5D VAS, SF-36, SGRQ und FACIT). Daher stellt sich die Frage, ob die vom IQWiG zitierten Übersichten repräsentativ für die regelmäßig in der Nutzenbewertung herangezogenen PRO- Instrumente sind.

Der G-BA hatte bei Veröffentlichung des neuen Methodenpapiers angeregt, bis auf Weiteres sowohl die bisher akzeptierten als auch die neuen 15 %-Schwellen parallel in den Dossiers darzustellen, um hieraus Erfahrungen zu deren Eignung abzuleiten.

Eine Übergangsphase von sechs Monaten ist sicherlich zu kurz, um einen umfangreichen methodischen Vergleich durchführen und bewerten zu können. Diese sinnvolle Erprobungsphase sollte verlängert werden, um die Übertragbarkeit der 15 %- Schwelle auf den AMNOG-Kontext weiter evaluieren zu können. Anschließend sollte eine Bewertung erfolgen, die widerspiegelt, ob eine generische Schwelle überhaupt sinnvoll ist und falls ja, wie hoch eine angemessene Schwelle ist. Andernfalls besteht die Gefahr, dass eine Grenze festgelegt wird, die potenziell patientenrelevante Unterschiede zwischen Therapien nicht mehr berücksichtigt.

Zusammenfassend schlägt die Novartis Pharma GmbH folgendes Vorgehen vor:

Gut validierte, etablierte und präspezifizierte Schwellen werden weiterhin akzeptiert und zentral veröffentlicht. Die bisherige Phase der parallelen Darstellung von Responseschwellen wird fortgesetzt (z. B. bis Ende 2022). Danach erfolgt eine wissenschaftliche vergleichende Evaluierung der Methodik und der Ergebnisse im Dialog mit allen an der Bewertung beteiligten Stakeholdern.

Verband Forschender Arzneimittelhersteller e. V.

Insgesamt ist festzuhalten, dass die Änderung der G-BA-Vorgaben für PRO-Responderanalysen nicht dem aktuellen Stand der wissenschaftlichen Erkenntnisse sowie den international anerkannten Kriterien und Standards der evidenzbasierten Medizin entspricht. Die Messlatte für bedeutsame patientenberichtete Ergebnisse wird damit zugleich substanziell erhöht, wobei patientengerechte Besonderheiten nicht berücksichtigt werden. Der Entwicklungsansatz der Wissenschaft, die Bewertungsstandards durch sinnvolle Qualitätskriterien zu verbessern, wird außer Acht gelassen. In Folge steht damit die Mehrzahl akzeptierter und international etablierter MID-Schwellen vor dem Aus. Der Nachweis von Verbesserungen, aber auch von Verschlechterungen kann damit erschwert sein. Der vfa schlägt daher vor:

1. Einzelfallprüfung statt „one-size-fits-all“

Der vfa hält eine bedenkenlose Anwendung des festgesetzten Richtmaßes für nicht angebracht. Die Eignung des Richtmaßes sollte in G-BA-Bewertungen auch weiterhin in jedem Einzelfall geprüft werden. Eine MID-Bewertung im festgelegten „one-size-fitsall“-
Ansatz kann bekannte Unterschiede der Patientensicht auf bedeutsame Ergebnisse nicht hinreichend berücksichtigen. Aus der Sicht des vfa sollte bei Änderungen der Methoden zur Beurteilung von MID die Wissenschaftlichkeit im Vordergrund stehen, nicht alleiniger Pragmatismus.

2. Allgemein akzeptierter Katalog von Bewertungskriterien

Die Bestrebungen der wissenschaftlichen Gemeinschaft um verbesserte Bewertungskriterien sollten nicht ignoriert, sondern unterstützt werden. Ziel sollte ein allgemein akzeptierter Katalog von Bewertungskriterien sein, der eine angemessene Beurteilung der Zuverlässigkeit von MID erlaubt. Dieser sollte auf Grundlage der bisherigen Empfehlungen und im weiteren gemeinsamen Dialog aus Wissenschaft, Patientenvertretern, Institutionen und Industrie entwickelt werden. Methoden zur Beurteilung von MID sollten erst dann geändert werden, wenn diese Diskussionen eine
ausreichende Einigkeit erreicht haben.

3. Verfahrenskonsistenz

Bis zur Festlegung einer verbesserten und allgemein akzeptierten Methode sollten die bisher geltenden Bewertungsmaßstäbe nicht geändert werden. Deshalb sollten alle bisher als etabliert bzw. validiert akzeptierten MID aus Gründen der Verfahrenskonsistenz auch weiterhin vom G-BA herangezogen werden.

Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF); Deutsche Gesellschaft für Pharmazeutische Medizin e. V.

Die DGPharMed empfiehlt dringend die zeitnahe Entwicklung einer verbindlichen, nationalen Leitlinie zu Patientenberichteten Endpunkten und entsprechenden Instrumenten/Skalen, auch um langfristig die Sicherheit und Planbarkeit für die Hersteller zu erhöhen. In die Erstellung dieser Leitlinie müssen außer dem IQWIG auch die wissenschaftlichen Fachkreise eingebunden werden. Neben den Mitgliedsgesellschaften der AWMF und einigen Lehrstühlen an deutschen Universitäten ist dies vor allem auch das CIPS (Collegium Internationale Psychiatriae Scalarum), das sich seit fast 50 Jahren mit klinischen Skalen zur Wirksamkeits-

und Verträglichkeitsbeurteilung von Interventionen in der psychiatrischen und psychopharmakologischen Forschung beschäftigt.

AbbVie Deutschland GmbH & Co. KG

Aus Sicht von AbbVie sollte die Festlegung einer generischen Responseschwelle entfallen, denn eine generelle Eignung sieht AbbVie basierend auf der Herleitung und Begründung des IQWiG nicht. Stattdessen sollte in einer Einzelfallüberprüfung anhand eines von Wissenschaftlern zuvor festgelegten Kriterienkatalogs und unter Einbezug von Patientenvertretern eine spezifische Responseschwelle festgelegt werden. Erst wenn sich auf diese Weise keine Responseschwelle ableiten lässt, kann ein pragmatischer Ansatz durch Wahl einer generischen Responseschwelle sinnvoll sein.

Deutschen Atemwegsliga e. V.

Die Umstellung der patientenberichteten Endpunkte auf eine Responseschwelle für Responderanalysen von mindestens 15 % der Skalenspannweite eines Instrumentes sollte solange ausgesetzt werden bis ausreichende wissenschaftliche Evidenz für diese Vorgehensweise vorliegt.

UCB Pharma GmbH

[Das vorgeschlagene Vorgehen ist] aus Sicht von UCB abzulehnen. Stattdessen muss die Patientenperspektive auch weiterhin im Rahmen der Nutzenbewertung berücksichtigt werden. Dafür sollten etablierte Responseschwellen (MIDs) weiterhin akzeptiert werden. Damit wäre auch eine Konsistenz zu der Zulassung gegeben. Ferner wäre eine Vergleichbarkeit zu früheren Nutzenbewertungsverfahren gewährleistet.

Des Weiteren sollte ein Kriterienkatalog zur Beurteilung der Validierung von MIDs im gemeinsamen Dialog von Industrie, Akademia, Behörden (IQWiG, BfArM, G-BA) auf der Basis von existierenden Arbeiten (z.B. Coon & Cook, 2008 [129], Devji et al., 2020 [130]) weiterentwickelt werden.

Astellas Pharma GmbH

Im Rahmen dieser Stellungnahme wird vorgeschlagen weiterhin validierte, allgemein akzeptierte MIDs in der Nutzenbewertung zu berücksichtigen und erstmal keine Änderungen der Modulvorlage vorzunehmen.

Stattdessen schlägt Astellas vor einen wissenschaftlich fundierten Kriterienkatalog zu etablieren, der regelhaft im Rahmen der Nutzenbewertung zur Beurteilung von Validierungsstudien anzuwenden ist. Dieser sollte nur im intensiven Austausch mit Vertretern aus Wissenschaft, Praxis, ggf. Zulassungsbehörde und pharmazeutischer Industrie entwickelt werden. Erst danach sollte eine Änderung der Modulvorlage angestrebt werden.

Bristol-Myers Squibb GmbH & Co. KGaA

- Keine allgemeine Schwelle gültig für alle Fragebögen, sondern Akzeptanz etablierter Schwellen, bis der wissenschaftliche Dialog abgeschlossen ist
- Fortführung der wissenschaftlichen Diskussion zur Bewertung der Responseschwellen und der Analysen der patientenberichteten Endpunkte für die Nutzenbewertung

129 Coon CD, Cook KF (2018): Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res; 27(1):33–40.

130 Devji T et. al (2020): Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ; 369:m1714. doi: 10.1136/bmj.m1714.

Merck Serono GmbH

Die vorgeschlagenen Änderungen zur MID-Bewertungsmethodik sollten gänzlich entfallen. Die Etablierung von MIDs sollte wissenschaftlich-methodisch hergeleitet werden und sich auch auf bisherige Erfahrungen/Praxis stützen.

Neue Standards zur MID-Bewertung sollten einem konsensbasierten, konstruktiven Dialog zwischen IQWiG, G-BA, Akademia und vfa folgen und schnellstmöglich gemeinsam erarbeitet werden. Das Ziel sollte sein, einen wissenschaftlich fundierten Kriterienkatalog zu etablieren, der für MIDs im Rahmen von Nutzenbewertungen regelhaft anzuwenden ist. Hierbei sollten u.a. auch die Ergebnisse der europäischen Innovative Medicines Initiative (IMI) Topic: „Establishing international standards in the analysis of patient reported outcomes and health-related quality of life data in cancer clinical trials“ Berücksichtigung finden. (IMI, 2019) 8

Biogen GmbH

Aus Sicht von Biogen sollten somit vorrangig jene MIDs in der Dossierbewertung Berücksichtigung finden, die für ein bestimmtes skalenbasiertes Instrument innerhalb der untersuchten Indikation und Patientenpopulation validiert sind.

Bundesverband der Pharmazeutischen Industrie e. V.

Im Stellungnahmeverfahren zum Methodenpapier des IQWiG wurden die neuen Schwellenwerte verschiedentlich kritisiert, insbesondere wurde auch deren mangelhafte wissenschaftliche Herleitung hervorgehoben.

Vor diesem Hintergrund sollten die überarbeiteten Dossiervorlagen, sofern der G-BA sich trotz der vorgebrachten Kritik am methodischen Ansatz des IQWiG orientieren will, methodisch so weit offengehalten werden, dass auch weiterhin von den Vorgaben des IQWiG abweichende Analysen vorgetragen werden können und der G-BA mithin eine Berichterstattung nicht auf die methodischen Vorgaben des IQWiG beschränkt. Dies ist zum Erhalt einer wissenschaftlichen Pluralität und zur Weiterentwicklung des methodischen Ansatzes insgesamt geboten. Hierzu gehört auch, dass der G-BA – ggf. auch abweichend vom IQWiG – Analysen, die ein abweichendes Analysemodell verwenden, hinsichtlich der Verwertbarkeit für die Zusatznutzenbewertung individuell inhaltlich prüft. Der G-BA sollte auch das bisherige Vorgehen bei der Bewertung der klinischen Relevanz beibehalten.

Deutsche Diabetes Gesellschaft e. V.

Die DDG plädiert für eine Überarbeitung des Beschlussentwurf in dem Sinne, dass die Festlegung der MID auf 15% der Skalenspannweite überdacht und individuell angepasst wird.

Pfizer Deutschland GmbH

Zusammengefasst ist Pfizer der Ansicht, dass die vorgesehenen Änderungen der Modulvorlage in der Anlage II.6 (Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen) bzgl. der PRO-Responderanalysen wissenschaftlich nicht hinreichend begründet und daher nicht gerechtfertigt sind. Eine mögliche Änderung der Vorgaben darf nicht zur Ablehnung von Responseschwellen führen, die bisher als validiert bzw. etabliert galten. Aus Sicht vom Pfizer sollte vor der Festlegung neuer Standards zur Bewertung von Relevanzschwellen der bisher stattgefundene Austausch mit Vertretern aus Wissenschaft, Praxis und pharmazeutischer Industrie intensiv fortgeführt werden. Offene Fragen sollten in einem wissenschaftlichen Diskurs erst geklärt werden, bevor eine solche Änderung mit potentiell bedeutsamen Auswirkungen auf die Nutzenbewertungspraxis vorgenommen wird. Pfizer ist hierbei anhaltend offen für einen Dialog.

Bis zur Festlegung einer verbesserten und allgemein akzeptierten Methode zur Bewertung von Responseschwellen sollten die bisher geltenden Bewertungsmaßstäbe nicht geändert werden. Deshalb sollten alle bisher als etabliert bzw. validiert akzeptierten MID aus Gründen der Verfahrenskonsistenz auch weiterhin vom G-BA herangezogen werden.

Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V.

[…] Die folgende Stellungnahme wurde von der ständigen Kommission Nutzenbewertung der AWMF verfasst. Die Deutsche Gesellschaft für Rheumatologie (DGRh), die Deutsche Gesellschaft für Pneumologie und Beatmungsmedizin (DGP) sowie die Deutsche Gesellschaft für Allergologie und klinische Immunologie (DGAKI) schließen sich der Stellungnahme an. […]

- Wenn der Grenzwert von 15% im Sinn einer Vereinheitlichung der Bewertungskriterien etabliert wird, müssen zusätzliche Ausnahmeregelungen festgelegt werden. Diese betreffen:

1. Besonderheiten des jeweiligen Krankheitsbildes sowohl bei der Schwere des Krankheitsbildes auch in Bezug auf die jeweils etablierten und standardisierten Skalen, siehe Stellungnahme DDG, DGIM und DGPPN
2. Größe der Studienpopulationen: Bei einer sehr kleinen Studienpopulation z. B. bei Arzneimitteln mit Orphan-Drug-Status aber auch bei molekular definierten Studienpopulationen ist eine größere Spannbreite erforderlich, um auch in diesen Arzneimittelbewertungen PRO berücksichtigen zu können.

- Eine „nationale Leitlinie“ unter Einbeziehung aller Stakeholder (HTA, Fachgesellschaften, Bundesoberbehörden u. a.) kann sicherstellen, dass die verschiedenen Bewertungs-instrumente für Zulassung, Nutzenbewertung und Leitlinien nicht differieren und zu Verunsicherung in der Versorgung führen.

IQVIA Commercial GmbH & Co. OHG

IQVIA schlägt vor, das bisherige Vorgehen bei der Bewertung der klinischen Relevanz beizubehalten, d.h. Responderanalysen auf der Basis eines präspezifizierten, validierten Responsekriteriums heranzuziehen. Für den Fall, dass ein etabliertes Kriterium nicht existiert, bietet der Vorschlag der Verwendung eines post-hoc Kriteriums von 15% der Skalenspannweite eine pragmatische Lösung und erscheint somit akzeptabel.

Unabhängig davon erhoffen wir uns die Aufnahme bzw. Weiterführung der Diskussion zu Bewertungsmaßstäben für Studien zur Herleitung von MIDs wie sie von der Gruppe um Gordan Guyatt angestoßen wurde.

Janssen-Cilag GmbH

Die Janssen-Cilag GmbH schlägt vor, die Thematik im Rahmen eines konstruktiven Dialogs zwischen IQWiG, G-BA und vfa unter Beteiligung von Vertretern der Akademie und der Entwickler der Erhebungsinstrumente zu vertiefen, um gemeinsam einen wissenschaftlich anerkannten Kriterienkatalog abzuleiten.

Bis zu einer Festlegung einer wissenschaftlich haltbaren und allgemein akzeptierten Methode für die Bestimmung der MCID sollten alle bisher als validiert und etabliert geltenden MCID beibehalten und aus Verfahrenskonsistenz weiterhin vom G-BA vorrangig in den Nutzenbewertungsverfahren herangezogen werden.

Bayer Vital GmbH

Das ursprüngliche Vorgehen sollte beibehalten werden. Für Fälle, in denen vorhandene Validierungsstudien zur anzusetzenden MID mit gravierenden Mängeln behaftet sind, kann erst nach intensivem methodischem Diskurs auf internationaler Ebene eine konsentierte

Alternative in Erwägung gezogen werden. Des Weiteren könnte in solchen Fällen anstelle von ankerbasierten MIDs auf verteilungsbasierte MIDs vorläufig ausgewichen werden.

Novo Nordisk Pharma GmbH

Novo Nordisk schlägt vor, von der Etablierung einer pauschalen Responseschwelle (MID) von 15 % in der Verfahrensordnung abzusehen.

Responderanalysen unter Verwendung einer wissenschaftlich und evidenzbasiert begründeten Responseschwelle (MID) sollten grundsätzlich indikationsspezifisch für die Nutzenbewertung herangezogen werden.

Zur Beurteilung der Stärke der Ergebnis(un)sicherheit sollte dann in einem zweiten Schritt die Validität der verwendeten Responseschwelle (MID) beurteilt werden.

Als Analysen höchster Aussagesicherheit ("Beleg") könnten dabei in Anlehnung an Revicki et al. 2008 Responderanalysen eingestuft werden, welche MIDs verwenden, die auf patienten-basierten und klinischen Ankern basieren.

Ecker + Ecker GmbH

Zusammenfassend erscheint uns der Nutzen des 15 %-Kriteriums fragwürdig, da den Vorteilen (Einfachheit) erhebliche Nachteile gegenüberstehen (u.a. methodische Fragwürdigkeit, Verlust an Sensitivität, unklare Anwendbarkeit auf alle Skalen). Zudem ist das Kriterium unnötig, da es dem G-BA schon jetzt freisteht, ungeeignete MCID abzulehnen.

Das Bemühen um eine einheitliche und einfache Regelung ist zu begrüßen. Dennoch halten wir es für sinnvoller, für die Zeit bis zum Vorliegen von akzeptierten Kriterien, nach denen die Qualität einer MCID beurteilt werden kann, bei der bisherigen Vorgehensweise zu bleiben. Nur so bleibt sichergestellt, dass der Bandbreite von Instrumenten und der Vielfalt von therapeutischen Kontexten Rechnung getragen wird und Rückschlüsse auf die Patientenrelevanz aus Ergebnissen komplexer Skalen getroffen werden können.

Deutsche Gesellschaft für Psychologische Schmerztherapie und -forschung e. V.

Die DGPSF hält die im Dokument „Tragende Gründe“ aufgeführte Begründung der Entscheidung für sachlich angemessen, da basierend auf einer sorgfältigen und qualitativ hochwertigen Analyse der vorhandenen Literatur. Vor diesem Hintergrund unterstützt die DGPSF die geplante Änderung der Modulvorlage in Anlage II (Frühe Nutzenbewertung).

Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V.; Deutsche Gesellschaft für Gerontologie und Geriatrie e. V.

Die DGGG stimmt der Präzisierung der Änderung im 5. Kapitel der Verfahrensordnung: Modulvorlage in der Anlage II (Frühe Nutzenbewertung) zu.

Deutsche Schmerzgesellschaft e. V.

Die Deutsche Schmerzgesellschaft e.V. hält die im Dokument „Tragende Gründe“ aufgeführte Begründung der Entscheidung für sachlich angemessen, da basierend auf einer sorgfältigen und qualitativ hochwertigen Analyse der vorhandenen Literatur.

Vor diesem Hintergrund unterstützt die Deutsche Schmerzgesellschaft e.V. die geplante Änderung der Modulvorlage in Anlage II (Frühe Nutzenbewertung).

Bewertung

Die Vorschläge zum weiteren Vorgehen waren im Wesentlichen: (1.) die geplanten Änderungen in der VerfO nicht zu übernehmen bzw. (2.) die Anwendung des Responsekriteriums von mindestens 15 % der Skalenspannweite nur für die

Fallkonstellationen vorzusehen, in denen keine etablierte/akzeptierte MID vorliegt bzw. (3.) die Fortführung der Diskussion und Entwicklung eines Kriterienkataloges zur Beurteilung der MID. Zudem wurde vorgeschlagen, alle MIDs um eine einheitliche „Sicherheitsspanne“ zu ergänzen und somit den diskutierten Unsicherheiten zu begegnen.

Diese genannten Vorschläge adressieren im überwiegenden Teil ein Fortführen der derzeitigen Praxis. Die Einwände der Stellungnehmenden haben insbesondere abgestellt auf die Auswirkungen der Anwendung einer Responseschwelle von 15 % auf verschiedene Fragebögen, das methodische Vorgehen zur Ableitung der Responseschwelle von 15 %, die aktuelle wissenschaftliche Diskussion zur Bewertung von MIDs sowie auf die Gegenüberstellung von Ergebnissen akzeptierter MIDs aus Beschlüssen des G-BA mit einer Responseschwelle von 15 %.

Insgesamt bleibt festzustellen, dass die Studien zur Bestimmung von MIDs in der Regel nicht dem aktuellen Stand der wissenschaftlichen Erkenntnis entsprechen. Dementsprechend gehen Responderanalysen auf Basis eines Responsekriteriums im Sinne einer MID mit wesentlichen Unsicherheiten bei der Interpretation der klinischen Relevanz des beobachteten Effektes einher.

Diese methodischen Unsicherheiten konnten von den Stellungnehmenden nicht ausgeräumt oder entkräftet werden. Zudem wurden keine substantiierten Alternativvorschläge unterbreitet, wie man eine zuverlässige Aussage zu einer klinisch relevanten Veränderung ableiten könnte.

Es zeigte sich ein allgemeiner Konsens, dass Responderanalysen allgemeine Vorteile gegenüber der Analyse stetiger Daten aufweisen.

Unbenommen der aktuellen und anhaltenden wissenschaftlichen Diskussion um Kriterien für die Entwicklung von einer MID und möglicher zukünftiger Standards ermöglicht die Anwendung eines Responsekriteriums von mindestens 15 % der Skalenspannweite zum jetzigen Zeitpunkt – im Gegensatz zu MIDs – die Berücksichtigung von aussagekräftigen Responderanalysen im Rahmen der Nutzenbewertung. Ein Wert von 15 % der Skalenspannweite soll hierbei sicherstellen, dass eine für die Patientinnen und Patienten hinreichend sicher spürbare Veränderungen abgebildet wird.

In Bezug auf das methodische Vorgehen zur Ableitung der Responseschwelle von 15 % ist festzuhalten, dass im IQWiG-Methodenpapier ein Wert von 15 % der Spannweite der jeweiligen Skalen als plausibler Schwellenwert für eine hinreichend sicher spürbare Veränderung empirisch gestützt hergeleitet und die Festlegung der Responseschwelle auf Basis des aktuellen Standes der wissenschaftlichen Erkenntnis vorgenommen wurde.

Hinsichtlich der Auswirkungen der Anwendung einer Responseschwelle von 15 % auf verschiedene Fragebögen ist festzustellen, dass die Definition der Responseschwelle von 15 % der Skalenspannweite dazu führt, dass zum Teil bisher verwendete Responsekriterien (MID) nicht mehr berücksichtigt werden. Dies steht der Anpassung der Anlage II.6 zum 5. Kapitel der VerfO jedoch nicht entgegen, da die Studien zur Bestimmung von MIDs in der Regel nicht dem aktuellen Stand der wissenschaftlichen Erkenntnis entsprechen. Auch kann nicht – insbesondere unter Berücksichtigung der Beispiele in der Nutzenbewertung seit Einführung des neuen methodischen Vorgehens – abgeleitet werden, dass der neu vorgeschlagene Wert von 15 % der Spannweite der jeweiligen Skalen den Nachweis von Vor- und Nachteilen von Therapien in patientenberichteten Endpunkten in relevantem Ausmaß erschwert. Es ist davon auszugehen, dass der Nachweis eines Effektes nicht von einer spezifischen MID abhängt (deren Validierung zudem aktuell in der Regel nicht dem aktuellen Stand der wissenschaftlichen Erkenntnis entspricht), sondern, dass ein Effekt auch mit dem neu vorgeschlagenen Wert von 15 % der Spannweite der jeweiligen Skalen nachweisbar ist.

Von den Stellungnehmenden wurde keine Gegenüberstellung von Ergebnissen akzeptierter MIDs aus Beschlüssen des G-BA mit einer Responseschwelle von 15 % vorgenommen.

Stattdessen wurde von den Stellungnehmenden auf Simulationen abgestellt. Aus der Simulationsstudie kann jedoch nicht abgeleitet werden, dass der neu vorgeschlagene Wert von 15 % der Spannweite der jeweiligen Skalen den Nachweis von Vor- und Nachteilen von Therapien in patientenberichteten Endpunkten erschwert.

In der Gesamtschau der in das Stellungnahmeverfahren eingebrachten Argumente konnten keine tragfähigen Alternativen gegenüber der Anpassung der Anlage II.6 zum 5. Kapitel der VerfO identifiziert werden.

Eine wie von den Stellungnehmenden gewünschte Fortführung der Diskussion um die Entwicklung eines Kriterienkataloges zur Beurteilung der MID findet aktuell und auch zukünftig, z.B. im SISAQOL-Projekt (Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data-Consortium), statt.

Mündliche Anhörung

gemäß § 35a Absatz 3 Satz 2 i.V.m. § 92 Absatz 3a SGB V i.V.m. § 91 Absatz 4 Nummer 1 SGB V des

Gemeinsamen Bundesausschusses

hier: Änderung der Modulvorlage

Videokonferenz des Gemeinsamen Bundesausschusses in Berlin

am 28. September 2021

von 10:00 Uhr bis 11:53 Uhr

– Stenografisches Wortprotokoll –

Angemeldeter Teilnehmender der Firma **UCB Pharma GmbH:**

Herr Andreas

Angemeldete Teilnehmende der Firma **Amgen GmbH:**

Frau Stein

Frau Lebioda

Angemeldete Teilnehmende der Firma **Novo Nordisk Pharma GmbH:**

Herr Dr. Bauer

Herr Dr. Kiencke

Angemeldeter Teilnehmender für den **Bundesverband der Pharmazeutischen Industrie e. V. (BPI):**

Herr Dr. Wilken

Angemeldete Teilnehmende der **Deutschen Atemwegsliga e. V.:**

Herr Dr. Kardos

Herr Prof. Dr. Worth

Angemeldeter Teilnehmender für die **Deutsche Gesellschaft für Innere Medizin e. V. (DGIM):**

Herr Prof. Dr. Sauerbruch

Angemeldete Teilnehmende der Firma **CIPS, Dr. Lorkowski:**

Her Prof. Dr. Weyer

Herr Prof. Dr. Görtelmeyer

Angemeldete Teilnehmende der Firma **Bristol-Myers Squibb GmbH & Co. KGaA:**

Frau Dr. Kupas

Angemeldete Teilnehmende der Firma **AbbVie Deutschland GmbH & Co. KG:**

Frau Dr. Sternberg

Herr Gossens

Angemeldete Teilnehmende der Firma **Roche Pharma AG:**

Herr Dr. Csintalan

Herr Dr. Knoerzer

Angemeldete Teilnehmende der Firma **Novartis Pharma GmbH:**

Frau Dr. Marx

Frau Dr. Eichinger

Angemeldete Teilnehmende der Firma **Astellas Pharma GmbH:**

Frau Zölch

Herr Dr. Groß-Langenhoff

Angemeldete Teilnehmende der Firma **IQVIA Commercial GmbH & Co. OHG:**

Frau Böhm

Angemeldete Teilnehmende der Firma **Pfizer Deutschland GmbH:**

Herr Dr. Miller

Herr Leverkus

Angemeldeter Teilnehmender der Firma **Bayer Vital GmbH:**

Herr Dr. Dintsios

Angemeldete Teilnehmende der Firma **GlaxoSmithKline GmbH & Co. KG:**

Herr PD Dr. Hennig

Herr Dr. Karl

Angemeldete Teilnehmende der Firma **MSD Sharp & Dohme GmbH:**

Frau Rettelbach

Herr Dr. Ziegler

Angemeldete Teilnehmende der Firma **Janssen-Cilag GmbH:**

Frau Dr. Huschens

Angemeldete Teilnehmende der Firma **Biogen GmbH:**

Frau Plesnila-Frank

Herr Dr. Dichter

Angemeldete Teilnehmende der Firma **Boehringer Ingelheim Pharma GmbH & Co. KG:**

Herr Pfarr

Herr Dr. Henschel

Angemeldeter Teilnehmender für den **Verband Forschender Arzneimittelhersteller e. V. (vfa):**

Herr Dr. Rasch

Angemeldete Teilnehmende der Firma **Merck Serono GmbH:**

Herr Schlichting

Frau Dr. Osowski

Angemeldete Teilnehmende der **Deutschen Gesellschaft für Hämatologie und Medizinische Onkologie e. V. (DGHO):**

Herr Prof. Dr. Wörmann

Angemeldete Teilnehmende für die **Deutsche Diabetes Gesellschaft e. V. (DDG):**

Herr Prof. Dr. Gallwitz

Herr Dr. Müller-Wieland

Beginn der Anhörung: 10:00 Uhr

Frau Dr. Behring (amt. Vorsitzende): Sehr geehrte Damen und Herren! Ich begrüße Sie ganz herzlich zu dem Unterausschuss Arzneimittel. Sie werden sich möglicherweise wundern, warum ich hier sitze und nicht Herr Hecken. Das liegt daran, dass Herr Hecken wie auch sein Stellvertreter, Herr Zahn, verhindert sind. Es wurde von der Ausnahmeregelung Gebrauch gemacht, die Sitzungsleitung auf die Geschäftsstelle zu übertragen, in diesem Fall auf mich. Mein Name ist Antje Behring. Ich leite die Abteilung Arzneimittel und führe somit heute durch diese Anhörung und die Sitzung des Unterausschusses.

Über diese Anhörung wird ein Wortprotokoll geführt. Somit sind alle Aussagen und Diskussionen unmittelbar festgehalten. Das bedeutet, ich bitte sie, vor Ihrem Wortbeitrag den Namen und die Institution zu sagen. Das macht es unseren Stenografen leichter, Ihre Stimme zuzuordnen. Ansonsten werden alle Diskussionsbeiträge Herrn Hecken dargestellt und vorgestellt, sodass der Informationsfluss gewährleistet ist.

Es geht heute um ein relativ ungewöhnliches Verfahren. Im Plenum wurde beschlossen, dass zu einer Änderung der Verfahrensordnung Stellung genommen werden kann. Das ist nicht üblich. Deswegen wurde das Stellungnahmerecht eingeräumt. Am 17. Juni wurde das beschlossen und das Stellungnahmeverfahren eröffnet. Ende Juli lief die Frist zur schriftlichen Stellungnahme zu dem Vorschlag zur Änderung der Modulvorlage aus. Zur Änderung der Darstellung der Ergebnisse der Responseschwelle haben Stellung genommen UCB Pharma, Amgen, Novo Nordisk, Merck Serono, der Bundesverband der Pharmazeutischen Industrie, die Deutsche Diabetes Gesellschaft, die Deutsche Atemwegsliga, die Deutsche Gesellschaft für Innere Medizin, CIPS Dr. Lorkowski, Bristol-Myers Squibb, AbbVie, Roche, Novartis Pharma, der vfa, Astellas, IQVIA, Ecker + Ecker, die Deutsche Gesellschaft für Psychiatrie und Psychotherapie, Psychosomatik und Nervenheilkunde, die AWMF, namentlich Herr Professor Dr. Wörmann, Pfizer, Bayer, GlaxoSmithKline, MSD, die Deutsche Gesellschaft für Psychologische Schmerztherapie und Forschung, die Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften, die AWMF, separat, Janssen-Cilag, die Deutsche Schmerzgesellschaft, Biogen sowie Boehringer Ingelheim.

Von den eingegangenen Stellungnehmern sind heute vertreten von UCB Pharma Herr Andreas, von Amgen Frau Stein und Frau Lebioda, von Novo Nordisk Herr Dr. Bauer und Herr Dr. Kiencke, vom BPI Herr Dr. Wilken, von der Deutschen Atemwegsliga Herr Dr. Kardos und Herr Professor Dr. Worth, von der Deutschen Gesellschaft für Innere Medizin Herr Professor Dr. Sauerbruch, von CIPS Herr Professor Dr. Weyer – Herr Professor Dr. Görtelmeyer von CIPS ist nicht zugeschaltet –, von Bristol-Myers Squibb Frau Dr. Kupas, von AbbVie Frau Dr. Sternberg und Herr Gossens, von Roche Herr Dr. Csintalan und Herr Dr. Knoerzer, von Novartis Frau Dr. Marx und Frau Dr. Eichinger, von Astellas Frau Zölch und Herr Dr. Groß-Langenhoff, von IQVIA Frau Böhm, von Pfizer Herr Dr. Miller und Herr Leverkus, von Bayer Herr Dr. Dintsios, von GlaxoSmithKline Herr Dr. Hennig und Herr Dr. Karl, von MSD Frau Rettelbach und Herr Dr. Ziegler, von Janssen-Cilag Frau Dr. Huschens, von Biogen Frau Plesnila-Frank und Herr Dr. Dichter, von Boehringer Herr Pfarr und Herr Dr. Henschel, vom vfa Herr Dr. Rasch, von Merck Serono Herr Schlichting und Frau Dr. Osowski. Herr Professor Dr. Wörmann von der DGHO hat abgesagt. Von der Deutschen Diabetes Gesellschaft sind noch zugeschaltet Herr Professor Dr. Gallwitz und Herr Dr. Müller-Wieland.

Ich danke Ihnen für die zahlreichen Stellungnahmen. Wir hatten mit dem Anschreiben insbesondere die pharmazeutischen Unternehmen aufgefordert, ob es nicht möglich wäre, eine Gegenüberstellung der Daten von Patient Reported Outcomes oder von komplexen Skalen zu liefern, in denen die 15-Prozent-Schwelle überschritten ist, und einmal nach der alten MID. Einige pharmazeutische Unternehmen haben uns darauf aufmerksam gemacht, dass sie Daten dazu schon eingereicht haben. Genau diese Daten haben wir von den Dossiers

gesehen, bei denen das tatsächlich vorlag, nämlich zwei verschiedene Ergebnisdarstellungen. Es haben uns aber keine Stellungnahmen erreicht, die ihre eigenen Ergebnisse gegenüberstellen. Anstelle dessen wurden uns Simulationsanalysen übermittelt, die dargestellt haben: Die bislang akzeptierten MIDs waren zum großen Teil unter 15 Prozent.

Diese Analyse wurde vom vfa eingereicht. Vielleicht können wir für die, die diese Analyse nicht kennen oder nicht in der Arbeitsgruppe sind, damit anfangen, in der mündlichen Anhörung darzustellen: Was war Inhalt der Simulationsanalyse, was war die Kernaussage? Mag sich jemand berufen fühlen, zu beginnen, jemand vom vfa oder jemand, der an dieser Arbeitsgruppe teilgenommen hat? – Herr Rasch meldet sich.

Herr Dr. Rasch (vfa): Ich kann gerne die Einleitung übernehmen. Dann kann jemand aus der Arbeitsgruppe zu den Details etwas sagen. Ich glaube, das würde auf jeden Fall Sinn machen.

Ich darf ganz kurz etwas zu den Kernpunkten unserer Stellungnahme sagen. Wir haben das Verfahren von Beginn an eng begleitet. Wir haben darauf hingewiesen, dass Konsens ist, dass MIDs üblicherweise mittels Validierungsstudien und mit Beteiligung von Patienten bestimmt werden – das wurde bei der letzten „IQWiG im Dialog“-Veranstaltung bestätigt –, um die verschiedenen Eigenschaften und die Besonderheiten der Betroffenen individuell zu berücksichtigen. Deswegen haben wir von Beginn an das generische Responsekriterium als Einheitswert für alle Patienten ganz grundsätzlich abgelehnt.

Wir haben auch festgestellt, dass die Festlegung, so wie sie erfolgt ist, nicht wirklich nachvollziehbar ist. Es soll gar keine Bewertung mehr für die Validierungsstudien stattfinden, sondern die Anwendung des Einheitsmaßes. Wie gesagt, wir konnten die arbiträre Festlegung nicht nachvollziehen. Was aus der Recherche des Methodenpapiers klar hervorgeht, ist, dass eine Einheitsschwelle kein plausibler und etablierter Ansatz ist. Das hat man so erkennen können.

Aber trotz dieser grundsätzlichen Ablehnung haben wir einige Sachen dennoch gemacht. Wir haben sie schon vor dem Stellungnahmeverfahren angefertigt. Das ist zum einen die angesprochene empirische Analyse aller MIDs in der Nutzenbewertung. Frau Behring, ich glaube, das ist das, was Sie als Erstes angesprochen haben. Wir haben gezeigt, dass die Mehrzahl der Schwellen deutlich unterhalb der vorgeschlagenen generischen Schwelle liegt. Wir haben auch gezeigt, dass trotz der Anwendung der neuen Schwelle in den laufenden Verfahren, was mehr oder weniger probehalber gemacht wurde, schon einige Verfahren aufgetaucht sind. Wir haben Beispiele in dem Stellungnahmeverfahren genannt. Inzwischen sind einige weitere dazugekommen, wo man erkennt, dass man mit der etablierten Schwelle einen Effekt sieht, mit der 15er-Schwelle diesen Effekt nicht mehr sieht. Vor allem haben wir noch etwas anderes gemacht. Wir haben seitens einer Arbeitsgruppe in der Industrie diese Simulation angefertigt, die für alle Beteiligten frei verfügbar ist. Sie ist allen offen zugänglich. Sie steht als Simulationstool zur Verfügung. Nach unserer Auffassung ist das genau das Tool, das man an dieser Stelle sehr gut gebrauchen kann, um solche Effekte der neuen Methodik zu bemessen; denn die Analyse einzelner Verfahren – wir haben die natürlich als Beispiele eingebracht – wäre nach unserer Auffassung nur ein Ausschnitt aus der Fülle dessen, was vorkommen kann. Sie würde die Eignung des Einheitsmaßes nicht sinnvoll untersuchen können. Wenn man sich einzelne Verfahren anschaut mit großen oder kleinen oder gar keinen Effekten, ist es letztlich egal, welches Maß man nimmt. Aber gerade in der Mitte stellt sich die Frage. Wir haben die Simulation eingebracht, um dem G-BA die Möglichkeit zu geben, die neue Methodik zu untersuchen. Das ist der Grund, warum wir das so gemacht haben.

Zu den Inhalten der Simulation kann gerne jemand etwas sagen, der aktiv daran beteiligt war.

Frau Dr. Behring (amt. Vorsitzende): Ich würde nun fragen, wer das ergänzen möchte. Herr Kardos, Sie haben sich gemeldet. Ich möchte aber zuerst dieses Thema abschließen und die klinische Stellungnahme am Ende aufgreifen. – Möchte jemand inhaltlich ergänzen? – Vorerst sehe ich keine Wortmeldung. Dann würde ich Frau Wieseler bitten, ihre Frage zu stellen.

Frau Dr. Wieseler: Vielen Dank. – Herr Rasch, Sie haben einen ganzen Strauß von Themen aufgemacht; zu den einzelnen Punkten könnte man jeweils etwas sagen. Ich werde mich jedoch auf die Frage von Frau Behring konzentrieren. Es geht um die Simulationsstudie im Vergleich zu einer Empirie aus den Daten der Nutzenbewertung. Sie haben mit dieser Simulationsstudie untersucht, was der Einfluss einer Responseschwelle auf die Power von Vergleichen ist. Sie haben festgestellt, dass abhängig davon, wo die Responderanteile liegen, es zu einem Powerverlust oder einem Powergewinn kommt. Das ist ganz kurz zusammengefasst das Ergebnis Ihrer Simulationsstudie.

Aus meiner Sicht hilft das nicht, abschließend die Frage zu beantworten, welchen Einfluss die 15-Prozent-Schwelle auf das Verfahren hat. Wir hätten tatsächlich gerne die Empirie gesehen. Ich denke, das wäre insgesamt für die Diskussion hilfreich. Denn die Frage ist: Welche Auswirkungen haben die Dinge, die Sie theoretisch simulieren, im konkreten Verfahren? In wie vielen Fällen kommt es tatsächlich zu Änderungen der Effekte? Welche Auswirkungen haben die Änderungen auf die Aussagen zum Zusatznutzen? Diese Fragen können wir nicht beantworten, weil Sie die entsprechende Empirie nicht vorgelegt haben. Sie haben in verschiedenen Stellungnahmen davon gesprochen – Sie haben das jetzt wieder aufgeführt –, dass diese Empirie selektiv wäre. Das kann ich nicht nachvollziehen. Die Empirie, die Sie bisher eingereicht haben, ist selektiv; das ist richtig. Wenn Sie aber entsprechend der Anforderungen des G-BA die Responderanalysen für alle Verfahren gegenübergestellt hätten, hätten wir keine selektive Darstellung, sondern eine Vollerhebung. Dann könnten wir beurteilen, was dieses Verfahren für die Beschreibung der Effekte bedeutet und in der Folge für Aussagen zum Zusatznutzen. Wir könnten verstehen, ob die Powerverschiebungen, die Sie in Ihrer Simulation zeigen, in der Praxis irgendwelche Auswirkungen haben oder auch nicht. Das können wir jetzt leider nicht machen, weil Sie die Daten nicht eingereicht haben, was ich bedauerlich finde, weil wir um diese Daten schon während der Diskussion des Methodenpapiers des IQWiG mehrfach gebeten haben.

Frau Dr. Behring (amt. Vorsitzende): Vielen Dank. – Herr Leverkus, Sie hatten sich gemeldet.

Herr Leverkus (Pfizer): Statistik ist im Prinzip nicht nur, dass man eine Formel anwendet und ausrechnet, Statistik ist eine Wissenschaft. Einige von uns haben das studiert. Es gibt zwei Arten von Forschung, die gemacht werden. Die eine Art ist im Prinzip, dass man mathematische Beweise macht, deren Verfahren irgendwie funktionieren. Die andere Art der Forschung sind Simulationsstudien. Das heißt, wenn ich wissen will, welchen Impact ein Verfahren hat, setze ich eine Simulation auf und lasse einen Parameterraum variieren. Das haben wir gemacht. Man kann sehen, bei welchen Parametern das funktioniert oder bei welchen Parametern das nicht funktioniert, wo man ein bisschen aufpassen muss, wo man diskutieren muss, was im Prinzip richtig ist. – Punkt eins.

Punkt zwei. Die Verfahren, die bis jetzt durchgeführt worden sind, sind Stichprobenverfahren. Die Verfahren in der Zukunft können ... [Tonausfall] bei einer sogenannten historischen Kontrolle. Das heißt, Sie wissen nicht, ob die Verfahren in der Zukunft genau den gleichen Parameterraum haben. Von daher ist das Verfahren, das wir hier vorgelegt haben, das, was man als Statistik-Wissenschaftler macht. Man macht eine Simulation und lässt den Parameterraum sehen. Man kann sehen, auf welche Parameter das zutrifft oder nicht. – Das ganz allgemein, um einzuordnen, was wir gemacht haben. Das ist das, was man als Statistiker macht, um Fragen zu beantworten.

Ich möchte an einen meiner Kollegen oder Kolleginnen weitergeben, die besser auf die Ergebnisse eingehen können und das im Detail interpretieren können, wenn das gewünscht ist.

Frau Dr. Behring (amt. Vorsitzende): Es haben sich einige gemeldet. Schön wäre es trotzdem gewesen, beides zu haben, sowohl die Simulation als auch konkrete Beispiele. Wir haben seit der Änderung des Methodenpapiers durchaus Unternehmen gehabt, die beides dargelegt haben. Wir haben seit Ende September zwölf Verfahren gehabt, wo wir beides beurteilen

konnten. Wenn wir die Zahl der Verfahren erweitern könnten, wäre unser Erfahrungsschatz größer. – Herr Andreas, bitte dazu.

Herr Andreas (UCB): Ich kann Friedhelm nur zustimmen. Wir decken den ganzen Parameterraum ab. Das heißt, jedes zusätzliche Beispiel wäre nur die Realisierung einer Verteilung, von der wir die Wahrheit nicht kennen. Bei der Simulation geben wir die Parameter vor und wissen, wie die wahre Verteilung wäre oder wie die Wahrheit ist, und können damit abschätzen, was die Irrtumswahrscheinlichkeit ist. Genau das haben wir gemacht. Wir konnten zeigen, dass es in sehr vielen Fällen, insbesondere bei einer Schiefe der Ausgangsverteilung, der Baseline-Verteilung, zu einem relevanten Powerverlust kommt, der bis zu 20 Prozent beträgt. Das heißt, hier besteht das große Risiko, dass man durch die Veränderung einer Relevanzschwelle – wir reden nicht mehr über eine MID, sondern eine Relevanzschwelle – Effekte übersieht, und zwar in beiden Richtungen, sowohl für einen Nutzen, aber insbesondere auch für einen Schaden.

Die andere Richtung, von der Frau Wieseler sprach, es gibt auch Fälle, wo das andersherum ist: Das ist in der Tat so. Es ist aber weniger so. Aus meiner Sicht kann aber nie das Argument sein, dass man sagt: Ich schieße einmal links vorbei und einmal rechts vorbei, und in der Mitte ist die Ente tot. Es kann nicht sein, dass man sagt, das ist eine Schwelle, mit der wir weiter arbeiten wollen. – Vielleicht möchte noch jemand anderes einsteigen.

Frau Dr. Behring (amt. Vorsitzende): Interessant fand ich Ihre erste Aussage: Mit der Simulation finden wir die Wahrheit. Ist das richtig? Sagten Sie, dass man mit der Simulation den wahren Wert findet?

Herr Andreas (UCB): Mit der Simulation deckt man den ganzen Raum ab, was Leverkus eben auch sagte. Wir nehmen nicht einzelne Beispiele, wo wir letztendlich nicht wissen, welche wahre Verteilung im Hintergrund liegt. Vielmehr ist das die Realisation eines Zufallsexperiments, das gewisse Schätzer zeigt, aber wir wissen nicht, wie letztendlich die Wahrheit aussehen wird. Dagegen geben wir im Rahmen der Simulation die Parameter vor. Sie sind transparent und damit für jeden nachvollziehbar.

Frau Dr. Behring (amt. Vorsitzende): Frau Wieseler, bitte.

Frau Dr. Wieseler: Vielen Dank. – Herr Andreas, Sie haben an verschiedenen Stellen gesagt, zum Beispiel kommt es zu einem hohen Powerverlust, wenn wir eine schiefe Baseline-Verteilung haben. Das kann ich nachvollziehen. Sie decken den ganzen Raum der Möglichkeiten ab und finden bestimmte Konstellationen, wo Sie mehr oder weniger hohe Powerverluste oder Powergewinne haben. Welche Situationen davon realistisch sind, in denen Daten vorkommen, häufig oder weniger häufig, das wissen wir aus der Simulation nicht. Deshalb ist die Empirie – ich rede nicht von Selektion, sondern von einer Vollerhebung – wichtig, um einschätzen zu können, welche Teile Ihrer Simulation für das Verfahren relevant sind. Ich weiß nicht, in wie vielen Dossiers und patientenrelevanten Endpunkten wir schiefe Baseline-Verteilungen haben. Wir sollten also die Empirie mit der theoretischen Simulation, die Sie vornehmen, kombinieren.

Frau Dr. Behring (amt. Vorsitzende): Herr Andreas, bitte.

Herr Andreas (UCB): Frau Wieseler, Sie sprechen von einer Vollerhebung. Von einer Vollerhebung sind wir weit entfernt. Auch wenn wir noch 20 weitere Verfahren sammeln, sind es nur einzelne empirische Ergebnisse. Man hat dann mehr Beispiele, aber letztendlich sind wir von einer Vollerhebung weit entfernt. Die werden wir nie leisten können. Es sind einzelne Realisationen, es sind einzelne Verfahren, einzelne Zufallsexperimente. Dafür ist der Begriff Vollerhebung aus meiner Sicht nicht adäquat.

Frau Dr. Behring (amt. Vorsitzende): Herr Leverkus und Herr Knoerzer haben sich zu den Einlassungen von Frau Wieseler gemeldet. Herr Leverkus, bitte.

Herr Leverkus (Pfizer): Wir wissen nicht, ob das, was in der Vergangenheit an Verfahren gemacht worden ist, in der Zukunft auch kommt. Insofern ist Ihre empirische Analyse nichts anderes als eine historische Kontrolle, zu der Sie immer sagen: Vorsicht! Was wir vorgelegt haben, ist, dass wir über den gesamten Parameterraum gehen und sagen, in welchen Situationen das schwierig wird. Wir haben nicht gesagt: In der Vergangenheit gab es in 23 Prozent der Fälle Schwierigkeiten und bei 12 Prozent nicht. Das kann man jetzt sehen, wenn man sich die Verteilung anschaut, wo es möglicherweise einen Powerverlust gibt oder auch nicht. Das ist die Forschungsfrage, die sich hier stellt, das ist die Forschungsfrage, die beantwortet ist. Wir können dreieinhalb Stunden darüber reden, dass die Industrie nicht geliefert hat und böse war. Ich weiß nicht, ob uns das weiterbringt. Wir sollten lieber darüber reden: In welchen Situationen funktioniert es, und in welchen Situationen funktioniert es nicht, und was müssen wir im Prinzip tun?

Frau Dr. Behring (amt. Vorsitzende): Ich hatte den Eindruck, dass wir darüber gar nicht mehr geredet hatten. Das war nur eine ganz kurze Zeit. – Ergänzend, Herr Knoerzer.

Herr Dr. Knoerzer (Roche): Frau Wieseler, ich wollte in eine ganz ähnliche Kerbe hauen. Mit den Beobachtungen selber lernen wir über die Eigenschaften dieser Grenzen ganz wenig. Wir sind in der sehr komfortablen Situation, dass wir ein Tool haben, das für alle verfügbar ist, mit dem wir sehr viel mehr lernen können, weil es eher zufällig ist. Das haben Herr Andreas und Herr Leverkus gesagt. Ich wollte sagen: Die Überlegung, die Sie anstellen, hat ein weiteres Problem. Denn dann können Sie jedes Konfidenzintervall und jeden Test, den wir bis jetzt vorgelegt haben, auch hinterfragen und sagen: Jetzt will ich mir das alles anschauen. Der Punkt ist: Wir haben jetzt den vollständigen Parameterraum abgedeckt. Das ist eine wunderbare Gesprächsgrundlage. Das ist unser Punkt; es ist etwas Proaktives.

Frau Dr. Behring (amt. Vorsitzende): Danke. – Frau Sternberg von AbbVie.

Frau Dr. Sternberg (AbbVie): Zum Thema Vollerhebung hat sich Herr Andreas schon geäußert. Ich wollte noch erwähnen: Wenn wir von einer Vollerhebung sprechen, die wir theoretisch hätten liefern können/sollen – wie auch immer –, müssen wir da nicht in Betracht ziehen, dass, wenn wir MIDs betrachten, diese populationsspezifisch sind, indikationsspezifisch sind? Alles, was wir liefern können, wäre populations- und indikationsspezifisch. So viele Beispiele für eine entsprechende Indikation und Population können wir gar nicht liefern, dass wir irgendwie in die Nähe einer Vollerhebung kommen. Von daher plädiere ich für die Wichtigkeit dieser Simulation, einen generellen Ansatz zu finden und sich nicht in Beispiele zu vertiefen. Wir haben jetzt Beispiele gesehen, bei denen es funktioniert hat. Wir haben auch Beispiele gesehen, bei denen wir offensichtlich einen Powerverlust hatten oder bei denen die Ergebnisse nicht konsistent waren. Von daher von meiner Seite das Plädoyer für die Simulationsanalyse und gegen eine Vollerhebung, die in keiner Weise eine Vollerhebung sein kann.

Frau Dr. Behring (amt. Vorsitzende): Frau Kupas, hat sich Ihre Meldung erübrigt, oder haben Sie noch etwas zu ergänzen?

Frau Dr. Kupas (BMS): Eigentlich wollte ich genau dasselbe sagen. Ich wollte etwas zur Vollerhebung sagen. Es ist keine Vollerhebung, sondern eine zufällige Stichprobe, was die alten Verfahren betrifft. Wir wissen überhaupt nicht, in welchem Parameterraum die sich befinden, wir wissen nicht, welchen wir in der Zukunft haben. Genau das haben die Vorredner auch schon gesagt.

Frau Dr. Behring (amt. Vorsitzende): Danke. – Frau Böhm.

Frau Böhm (IQVIA): Mein Beitrag hat sich erübrigt.

Frau Dr. Behring (amt. Vorsitzende): Danke. – Herr Rasch.

Herr Dr. Rasch (vfa): Ich möchte ganz kurz auf das antworten, was Frau Wieseler gesagt hat, die Anforderung, eine Vollerhebung durchzuführen. Man muss ganz grundsätzlich sagen, dass

die Festlegung der Methoden für die Nutzenbewertung und die Überprüfung der Eignung dieser Methoden beim IQWiG liegt. Das heißt, hier muss gewährleistet sein, dass die Methodik, die vorgeschlagen wird, tatsächlich den Standards der evidenzbasierten Medizin entspricht. Jetzt die Forderung nach einer Vollerhebung der Daten für die letzten zehn bis elf Jahre aufzustellen, ist in gewisser Weise eine Umkehr der Beweislast, die nach unserer Auffassung nicht wirklich tragbar ist, in der Art: Beweisen Sie mit einer Vollanalyse, dass unsere vorgeschlagene Einheitsschwelle nicht passt. Ich glaube, das funktioniert so nicht. Die Empirie und Simulationen hätten wir von Beginn an bei der Festlegung der Methoden erwartet. Aber unabhängig davon, ob Simulation, Vollerhebung, wie das alles zueinander steht: Zurück zum Anfang. Was in sehr vielen Stellungnahmen zu sehen war, war die Botschaft, dass eine Einheitsschwelle nicht angemessen ist, weil sie nicht wirklich plausibel hergeleitet wurde, aber viel mehr noch, weil das kein Stand der wissenschaftlichen Erkenntnisse ist, und sie berücksichtigt nicht die Besonderheiten der einzelnen Krankensituation und Patientensituation. Das ist aus meiner Sicht viel wichtiger als die Diskussion, was besser geeignet ist.

Frau Dr. Behring (amt. Vorsitzende): Herr Dintsios, bitte.

Herr Dr. Dintsios (Bayer): Ich war ein bisschen überrascht, dass Frau Wieseler – das war auch in der Anhörung beim IQWiG – Simulationen ad acta legt. Simulationen sind ein Mittel, das in der Statistik durchaus verwendet wird. Das IQWiG selber verwendet sie auch. Bei den binären Daten hat das IQWiG damals für seinen endpunktbestimmten Zusatznutzen – Matrix nenne ich es der Einfachheit halber – auch Simulationen verwendet. Erkenntnistheoretisch bringen sie uns sehr viel weiter, weil sie den ganzen Datenraum abdecken und nicht auf spezielle Einzelfälle zurückgreifen, die nie die gesamte Realität abbilden können, ob Sie wollen oder nicht. Ich gebe ein ganz kleines Beispiel aus der Gesundheitsökonomie, aus der ich primär stamme. Nehmen wir an, Sie haben eine Visuelle Analogskala, die von 0 bis 100 geht, und Sie haben eine Bevölkerung, die im Durchschnitt bei 67 Jahren liegt. Da liegt die altersbedingte Basisnutzwertangabe bei 75 Prozent. Bei 100 liegt ein Neugeborenes, 0 bedeutet Tod. Das ist ein sehr einfaches Beispiel. Wenn sie dann auf der Skalenbreite um 15 Prozent nach oben bei einer Lebensqualitätsverbesserung ansetzen wollten, dann haben Sie ein Problem: Die erreichen Sie gar nicht mehr. Sie kommen in ganz andere Schwierigkeiten, weil Skalenbereiche, die das IQWiG ansetzt, in der Bevölkerung nicht mehr erreichbar sind. Das ist nur ein kleines Beispiel, wieso die extern gesetzten Schwellenwerte erkenntnistheoretisch in meinen Augen nicht wissenschaftlich sind. Sie sind der Versuch, eine gewisse Sicherheit einzubinden, was vielleicht in den Augen des IQWiG legitim erscheint, aber sie sind weit weg von pragmatischen Vorgängen. Sie werden auch von niemand anders auf der Welt so angesetzt und auch umgesetzt. Ich glaube, das müsste schon ein bisschen zu denken geben, auch dem IQWiG selber.

Frau Dr. Behring (amt. Vorsitzende): Vielen Dank, Herr Dintsios. – Ich glaube, das Ganze hatte auch eine Ursache: Die MIDs waren nicht ganz so valide, wie man sich das gewünscht hätte. – Frau Müller.

Frau Dr. Müller: Ich muss leider zurück zur Simulation, weil ich da eine ganz einfache Verständnisfrage habe. Ich bin keine Methodikerin, deshalb bitte ich zu entschuldigen, wenn es eine dumme Frage ist. Ich habe eine ganz grundsätzliche Sache bei der Simulation oder den Schlüssen, die Sie daraus gezogen haben, nicht verstanden. Ich konnte nachvollziehen, dass Sie an verschiedenen Parametern gedreht haben, um zum Beispiel festzustellen, was Einfluss hat, beispielsweise schiefe Baseline-Verteilung. Das kann ich verstehen. Auf die Frage, wie sich das in der Praxis auswirkt, ist Frau Wieseler eingegangen. Was ich nicht verstanden habe, was Sie alle verwendet haben – auch in der vfa-Stellungnahme und bei Pfizer wurde das aufgeführt –: Sie haben gesagt, die Power wird erniedrigt durch Verwendung des 15-Prozent-Schwellenwertes statt der bisherigen MID. Ich zitiere jetzt aus der vfa-Stellungnahme: Insbesondere kann damit der Fehler zweiter Art vorliegen, die Wahrscheinlichkeit, keinen Effekt anzuerkennen, obwohl ein Unterschied vorliegt. – Ein vorhandener Unterschied wird

also nicht erkannt, könnte damit untersucht werden. Da habe ich ein Verständnisproblem insofern, als wir hier eigentlich nicht über einen Unterschied reden, der statistisch vorhanden ist und vielleicht nicht detektiert wird, sondern, soweit ich das bisher verstanden habe, reden wir bei MIDs darüber: Was spürt ein Patient? Die ganzen komplexen methodischen Verfahren dienen dazu – so hatte ich das bisher verstanden –, dass man in einer bestimmten Erkrankungssituation sagen kann: Spürt der Patient einen Unterschied? Im Moment wird wissenschaftlich angezweifelt, ob das bisherige Vorgehen adäquat ist. Dann ist für mich unklar, wie man von einem existierenden Unterschied sprechen kann, der möglicherweise mit der generischen 15-Prozent-Schwelle nicht detektiert wird, wenn für mich die Frage ist: Was ist überhaupt ein Unterschied? Ich habe das nicht verstanden. Für mich geht die Diskussion eigentlich darum: Was ist ein Unterschied, der gespürt wird, und wie kann man dann sagen: „Es gibt einen Powerverlust“? Ein Unterschied wird nicht erkannt, wenn man gar nicht weiß, ob mit den bisher etablierten MIDs überhaupt ein relevanter Unterschied erfasst wird. – Ich weiß nicht, ob Sie verstehen, was ich meine. Ich verstehe nicht, wie man im Zusammenhang mit der Simulation davon sprechen kann, dass man die Häufigkeit der Fehler zweiter Art überhaupt untersuchen kann, wenn man keinen Wert hat, auf den man sich bezieht. Man kann nicht sagen, es ist ein Unterschied da, der nicht erkannt wird, wenn für mich die Diskussion darum geht: Was ist überhaupt ein Unterschied? – Vielleicht bin ich auch blöd und habe einen Knoten im Gehirn; das kann sein. Aber es wäre gut, wenn ich das verstehen würde, was die ganze Zeit diskutiert wurde.

Frau Dr. Behring (amt. Vorsitzende): Herr Andreas, bitte.

Herr Andreas (UCB): Frau Müller, wir müssen zwei relevante Unterschiede unterscheiden. Auf der einen Seite – da haben Sie recht – geht es um die MIDs, wo im Regelfall und bei den meisten etablierten publizierten MIDs die 15-Prozent-Regel eine Erhöhung der individuellen Patienten-MID bedeutet. Wenn wir von einer 100er-Skala ausgehen, geht es von 10 auf 15 Punkte hoch. Das IQWiG sagt auch: Häufig ist es 10. Die 15 ist dann ≥ einer spürbaren Schwelle. Was wir in der Simulation gemacht haben, ist: Wir haben zwei Behandlungen miteinander verglichen. Das heißt, hier geht es um einen statistischen Test. Wie sieht es aus, wenn wir von einem gewissen Behandlungseffekt ausgehen, dass die Behandlung A im Vergleich zur Behandlung B eine höhere Responserate bezüglich dieser festen MID hat, welche Power hat der Test, mit welcher Wahrscheinlichkeit kann ich den Unterschied zwischen den beiden Behandlungen zeigen? Hier hat sich gezeigt: Wenn ich bei gleichen Populationen, bei gleichem Behandlungseffekt die MID – was eine Response ist – von 10 auf 15 Prozent hochsetze, sinkt in vielen Fällen die Power. Andersherum gesagt: Ich bräuchte eine höhere Fallzahl, um einen statistisch signifikanten Unterschied zwischen den beiden Behandlungen zu zeigen. – Hilft Ihnen das?

Frau Dr. Behring (amt. Vorsitzende): Ergänzend dazu, Frau Kupas.

Frau Dr. Kupas (BMS): Frau Müller, Ihre Frage ist sehr berechtigt. Diese Frage kann man auch mit der empirischen Auswertung nicht lösen; denn auch da schauen wir nur: Gibt es einen Unterschied von 10 vs. 15?, aber nicht: Spürt der Patient das? In der Simulation haben wir angenommen, wir haben einen echten Effekt, und haben uns angeschaut: Wie verändert sich die Power, wenn ich die MID erhöhe? Das heißt, wir haben den echten Effekt angenommen, wir haben angenommen, dass der Patient es spürt, und haben dann festgestellt, dass die Power in einigen Fällen heruntergeht.

Frau Dr. Behring (amt. Vorsitzende): Frau Böhm, vielleicht ergänzend dazu.

Frau Böhm (IQVIA): Der Vorteil der Simulation ist, dass ich die Behandlungsarme so simuliere, dass ein tatsächlicher Effekt da ist. Wir haben geschaut, je nachdem, welche Schwelle wir angenommen haben, wie groß die Anteile der Responder sind, und haben die verglichen. In diesem Zusammenhang haben wir den Powerverlust in einigen der Konstellationen, die wir untersucht haben, gesehen.

Frau Dr. Behring (amt. Vorsitzende): Frau Müller.

Frau Dr. Müller: Was Frau Kupas eben gesagt hat, hat mir sehr geholfen. Ich habe es so verstanden, dass Sie als echten Effekt, auf den sich sowohl der Powerverlust als auch insbesondere die Frage „Wie häufig tritt ein Fehler zweiter Art auf?“, dass ein existierender Unterschied nicht detektiert wird, als echten Unterschied die bisher akzeptierte MID gesetzt haben.

(Herr Andreas (UCB): Ja!)

Die Diskussion geht ja darum, ob die bisherigen Validierungsmethoden für die MIDs adäquat sind. Der Fehler bezieht sich also auf etwas, was wissenschaftlich im Moment zur Diskussion steht, und nicht auf einen Unterschied, der gesetzt ist. – Dann habe ich das verstanden. Die anderen Punkte waren mir schon klar. Mir ist klar, dass man in solchen Fällen potenziell eine größere Fallzahl braucht, dass es vorkommen kann, wenn man eine Schwelle, wo man Response sieht, höher setzt. Die anderen Sachen habe ich verstanden.

Die Frage ist – das ist für mich ein wichtiger Punkt –, das heißt, dass sich die ganzen Berechnungen auf die bisher etablierten MIDs beziehen, das Gleiche, was bei dem empirischen Vergleich ein Problem ist, das man bloß anschaut, dann allerdings in der Praxis: Wo kommt etwas anderes heraus? Eine Aussage darüber, was näher am wahren Effekt ist, den man sehen will, erlaubt weder das eine noch das andere für mich. – Das ist jetzt meine Auffassung, so wie ich das verstanden habe.

Frau Dr. Behring (amt. Vorsitzende): Herr Skipka.

Herr Dr. Skipka: Vielen Dank. – Ich möchte versuchen zu erklären, warum es sinnvoll ist, neben der Simulation empirische Daten zu betrachten. Von verschiedenen Personen wurde richtigerweise gesagt, eine Simulation hat den Vorteil, dass man die Wahrheit kennt. Man spannt einen Parameterraum auf. Das waren eine Handvoll Parameter, die man variieren kann, und man kann sehen, was passiert. Das ist wunderbar. Was man in der Simulation gesehen hat, ist, dass abhängig von den Risiken man, wenn man von 10 auf 15 Prozent als Responsekriterium geht, einen Powergewinn oder auch einen Powerverlust hat. Das Ganze ist sogar ziemlich symmetrisch, wenn man sich das mathematisch anschaut.

Jetzt stellt sich aber für die Praxis die Frage: Wo bewege ich mich mit den Basisrisiken? Ich kann nicht einfach behaupten, in vielen Fällen habe ich einen Powerverlust, wenn ich auf 15 Prozent gehe. Man kann genauso viele Fälle finden, wo man einen Powergewinn in derselben Größenordnung hat. Das alles zeigt die Simulation. Die Frage ist jetzt in der Praxis: In welchen Arealen des Parameterraums bewegt man sich? Das kann man nur feststellen, indem man sich die Empirie anschaut. Insbesondere muss man schauen, wie hoch die Risiken sind. Sind sie unter 50 Prozent, sind sie in der Nähe der 50 Prozent, oder sind sie oberhalb von 50 Prozent? Das ist das Entscheidende, und das kann ich mir nur empirisch anschauen. Das kann mir die Simulation nicht sagen.

Frau Dr. Behring (amt. Vorsitzende): Danke. – Ergänzend, Frau Wieseler.

Frau Dr. Wieseler: Ich möchte den Begriff Vollerhebung klären, den ich verwendet habe. Ich meine natürlich nicht eine volle Erhebung auf alle potenziell denkbaren Konstellationen, sondern, so wie Herr Skipka das beschreibt, eine Beschreibung dessen, was wir in den Nutzenbewertungen sehen, um einordnen zu können, welche Auswirkung dieses Verfahren in den konkreten Nutzenbewertungen hat. Natürlich kann es sein, dass die Nutzenbewertungen in der Zukunft etwas anders ausfallen als in der Vergangenheit. Nichtsdestotrotz würde für die Diskussion aus den Gründen, die Herr Skipka beschrieben hat, die Empirie helfen, dann in der Tat nicht selektiv, sondern repräsentativ für die Nutzenbewertungsverfahren, die wir bereits gesehen haben.

Frau Dr. Behring (amt. Vorsitzende): Vielen Dank. – Frau Kupas, Sie wollten zu dem, was Frau Müller gesagt hat, noch etwas ergänzen oder richtigstellen?

Frau Dr. Kupas (BMS): Ich glaube, es ist ein bisschen schief angekommen, Frau Müller. Wir haben einen echten Unterschied zwischen den Verteilungen angenommen und haben uns dann angeschaut: Wie viel Power haben wir mit der MID von 10 vs. der MID von 15? Wir haben nicht operierend aus der MID von 10 den Unterschied simuliert, sondern haben einen echten Unterschied der beiden Verteilungen angenommen. Da kann man die Größe des Effektes variieren und kann sich in der Simulation anschauen: Wie verändert sich die Power zwischen den beiden MIDs?

Frau Dr. Behring (amt. Vorsitzende): Herr Miller, Sie hatten etwas direkt dazu?

Herr Dr. Miller (Pfizer): Das, was Frau Kupas eben gesagt hat, wollte ich auch zu Protokoll geben. Von daher hat sich mein Beitrag erübrigt.

Frau Dr. Behring (amt. Vorsitzende): Danke. – Herr Leverkus, zu dem, was gerade vom IQWiG gesagt worden ist.

Herr Leverkus (Pfizer): Man muss, wenn man eine Simulationsstudie hat, auch schauen, was für einen Einfluss das in der Praxis hat. Wir legen das zurzeit immer vor. Wir machen im Prinzip beides. – Punkt eins.

Punkt zwei. Die Baseline-Verteilungen liegen auch vor. Die sind im Dossier drin. Von daher verstehe ich nicht, warum man die Vollerhebung machen muss.

Wie gesagt, wir haben die Simulation vorgelegt. Dazu kann man etwas sagen. Wir können sicherlich in einige Verteilungen hineinschauen: Sind die linksschief oder nicht linksschief?

Frau Dr. Behring (amt. Vorsitzende): Herr Hennig, eine letzte Ergänzung zu dem methodischen Teil.

Herr Dr. Hennig (GSK): Ich möchte zu dem zurückkommen, was Herr Rasch gerade gesagt hat, nämlich auf den Ausgangspunkt, ob das MID-15-Kriterium geeignet ist, valide Unterschiede festzustellen. Neben den ganzen Limitationen, die Herr Rasch schon ins Protokoll gegeben hat, die ich nicht wiederholen möchte, geht es mir darum: Was hilft uns eine Empirie, so wie sie vom IQWiG gefordert wurde, bei der Beantwortung jener Frage? Am Ende einer Empirie würden Zufallsbefunde aus den Verfahren stehen, die abgeschlossen sind, aber es kann mit dieser Empirie kein Beweis erfolgen, dass die MID 15 ein geeignetes Maß ist. Ich wollte nur daran erinnern: Worum geht es grundsätzlich? Aus meiner Sicht geht es um die Frage, ob eine MID von 15, ob das vorgeschlagene Maß vom IQWiG überhaupt sinnvoll ist. Ich bezweifle sehr, dass mit Empirie ein Beweis in dieser Frage möglich ist. – Vielen Dank.

Frau Dr. Behring (amt. Vorsitzende): Direkt dazu, Frau Wieseler.

Frau Dr. Wieseler: Vielen Dank, Herr Hennig. Das gibt mir Gelegenheit, einen Punkt zu klären, der in verschiedenen Stellungnahmen vielleicht nicht richtig verstanden ist. Der Wert, den wir hier vorlegen, diese 15 Prozent, ist explizit keine MID, soll es auch nicht sein und kann es auch nicht sein, weil, wie in vielen Stellungnahmen richtig beschrieben wird, eine MID von ganz vielen Faktoren abhängig ist: von der Indikation, vom Schweregrad der Erkrankung, von der Richtung der Veränderung. Das wird in der Literatur umfangreich diskutiert. Aus diesem Grunde haben wir die Notwendigkeit gesehen, einen anderen Vorschlag zu machen. Aber die 15 Prozent sind explizit keine MID.

Die Frage, warum wir überhaupt auf der einen Seite simulieren und auf der anderen Seite eine Empirie benötigen, ist eigentlich eine Frage, die von der Industrie aufgeworfen wurde, weil die Behauptung in den Raum gestellt wurde, dass ich damit den Zusatznutzen nicht mehr zeigen kann. Da war unser Vorschlag: Wenn Sie annehmen, dass Sie mit der 15-Prozent-Schwelle den Zusatznutzen nicht mehr zeigen können, dann analysieren Sie doch bitte Ihre Daten, damit wir das empirisch beurteilen können. Daraufhin haben Sie die Simulation vorgelegt. Das ist auch ein Verfahren. Aber jetzt müssen wir schauen, wie viel von der Simulation in der Realität stattfindet. Dafür brauchen wir die Empirie. – Das war die Historie der Diskussion um die Simulation und die Empirie.

Frau Dr. Behring (amt. Vorsitzende): Ich würde gerne Herrn Skipka direkt dazu das Wort geben.

Herr Dr. Skipka: Vielen Dank. – Frau Wieseler hat es genau so gesagt, wie ich es sagen wollte. Ich ziehe zurück.

Frau Dr. Behring (amt. Vorsitzende): Frau Müller, bitte.

Frau Dr. Müller: Ich wollte genau das sagen, erläutern, wozu das dienen soll.

Vielleicht eine Bitte. Ich sage es etwas überspitzt: Auch wenn Sie die Simulation für die geeignetere Methode halten, die bestimmte Fragen und mögliche Problemfelder gut aufzeigen kann oder gleiche Verteilung, was spricht für Sie dagegen, zusätzlich, da Sie erfreulicherweise zum Großteil sowohl die 15 Prozent als auch die alten MIDs in den Dossiers liefern, also die Möglichkeit da ist – das sollte für einen gewissen überschaubaren Zeitraum vollständig sein und nicht selektiv, je nachdem, wo sich der Effekt zeigt; denn dann geht das Ganze nicht mehr –, das anzusehen, um die Behauptung, die aufgestellt wurde, zu verifizieren oder auch nicht, dass man in vielen Fällen bei den PROs überhaupt nichts mehr sehen würde, indem man die Relevanzschwelle statt der MIDs verwendet? Das wäre, auch wenn es nur eine Augenblicksaufnahme über einen begrenzten Zeitraum ist, etwas, was bei der Diskussion hilft. Was spricht dagegen, das zusätzlich vorzulegen? Das habe ich bisher noch nicht verstanden. Es ist eine Ergänzung, es macht nicht das Gleiche. Was spricht für Sie dagegen?

Frau Dr. Behring (amt. Vorsitzende): Herr Schlichting, bitte.

Herr Schlichting (Merck Serono): Der Ausgangspunkt ist, dass wir hier einen künstlichen Schwellenwert haben und jetzt nachträglich versuchen, das irgendwie zu rechtfertigen, das heißt, durch Empirie oder die Eigenschaften versuchen, mit Simulationsstudien klarzumachen. Die Empirie gibt uns nur Auskunft über die Vergangenheit. Aber es werden weitere Fragebögen entwickelt. Wir werden in der Zukunft mehr PROMIS-Fragebögen sehen, die zur Bewertung anstehen. Es ist total unklar, wie der künstliche, allumfassende 15-Prozent-Schwellenwert überhaupt umgesetzt werden soll. Vor diesem Hintergrund ist es wirklich sehr fragwürdig, ob ein solcher künstlicher Schwellenwert überhaupt für die Nutzenbewertung sinnvoll ist. Das ist die entscheidende Frage. Ich würde dazu gerne die Kliniker hören. – Danke.

Frau Dr. Behring (amt. Vorsitzende): Herr Dintsios, bitte.

Herr Dr. Dintsios (Bayer): Es sind zwei Punkte, die ich anmerken will. Der eine ist vom Vorredner schon angerissen worden. Ob ich mit meinem künstlichen Schwellenwert von 10 auf 15 Prozent gehe oder von 7 auf 15 Prozent oder von 3 auf 15 Prozent – bei den Nutzwerten gibt es geankerte Differenzen, die bei 3 liegen –, über alles hinweg im Sinne eines Gießkannenverfahrens, diese Schwelle, die Methode, die aus zwei Literaturquellen stammt, beim finalen Methodenpapier 6.0 noch ein paar dazugenommen wurden und einen Mittelwert für den … [akustisch nicht zu verstehen] darstellte, ist in meinen Augen kein Erkenntnisgewinn. Ich kann das IQWiG in einer Sache verstehen: wenn es behauptet, die MID-Studien, die Validierungsstudien aus alten Zeiten sind, qualitativ betrachtet, nicht das Optimum. Da gehe ich durchaus mit; da habe ich überhaupt kein Problem. Wenn man einen Sicherheitsabstand einbauen wollte, eine Art Sicherheitsmarge, warum nicht auf die jeweiligen MIDs? Wieso einen 15-Prozent-Schwellenwert anlegen, der dann, wenn Sie sich die Heterogenität der MIDs anschauen, für einige dieser MIDs eine Verdreifachung darstellt? Diese Antwort bleibt das IQWiG bis heute schuldig. Ich habe mich damals in meiner ersten für Bayer formulierten Stellungnahme beim IQWiG erkundigt. Da kam nichts anderes als: Es kommt eine Setzung. Die Setzung, die extern ist, mag aus Sicht des IQWiG durchaus rechtens sein, aber es ist rein wissenschaftlich nicht unbedingt abgeleitet, nicht sauber abgeleitet. Sie ist auch nicht so leicht erzielbar. Ich spreche für mein Unternehmen. Ich bin ein Gegner davon. Ich habe keine Lust, alte Daten auszuwerten, bei denen eine Nutzenaussage, gestützt auf eine alte MID, getätigt wurde, und dann in eine Diskussion zu gehen, wenn es eine Änderung gibt,

einen Zusatznutzen theoretisch ableiten oder genau das Gegenteil. Ich glaube, das versteht jeder von uns.

Frau Dr. Behring (amt. Vorsitzende): Herr Dintsios – –

Herr Dr. Dintsios (Bayer): Lassen Sie mich ganz kurz ausreden. Ich habe den Satz noch nicht beendet. – Die Daten von allen Unternehmen in einen Pool hinzuzufügen, ist auch grenzwertig, auch kartellrechtlich. Aber das ist dem IQWiG anscheinend egal.

Frau Dr. Behring (amt. Vorsitzende): Herr Dintsios, ich glaube, Sie brauchen die ganze Diskussion, die schon einmal geführt worden ist, nicht zu wiederholen. Das ist in einem anderen Kontext.

(Herr Dr. Dintsios (Bayer): Ich glaube, Sie waren damals nicht dabei, Frau Behring, oder waren Sie damals dabei? Da waren Sie nicht dabei!)

– Ich weiß nicht, was das gerade zur Sache tun soll. – Frau Teupen, Sie haben das Wort zu einer neuen Frage.

Frau Teupen: Einige Stellungnehmer haben sich sehr konkret zum SF-36 und SF-12 geäußert und das Problem mit den zwei Dimensionen, die wir haben, angesprochen. Das war einmal von Novo Nordisk, GSK und Janssen. Die zitieren eine Publikation von Bjorner von 2011. Vielleicht können Sie das erläutern, weil das relativ pragmatisch erschien, wo spezifisch bei dem SF-36-Fragebogen ein Problem herrschen könnte.

Frau Dr. Behring (amt. Vorsitzende): Die Firmen Janssen, Novo Nordisk und GSK wurden angesprochen. Es ging konkret um den Fragebogen SF-36. – Frau Huschens, bitte.

Frau Dr. Huschens (Janssen-Cilag): Wir hatten in unserer Stellungnahme zum SF-36 geschrieben. In einer Studie war der SF-36 enthalten. Wir sahen uns vor der Situation, dass bei dem SF-36 die Populationen auf ein bestimmtes Jahr normiert sind und dass es je nach Verwendung des Jahres unterschiedliche Skalenspannweiten gibt. Es ist ein Unterschied, ob ich den SF-36 auf das Jahr 2009 auf die US-Population normiert habe oder auf die US-Population von 1998. Das ergibt andere Skalenspannweiten.

Der zweite Punkt war, dass bei der Bildung der Summenscores, beim MCS und PCS, die entsprechenden acht Komponenten eingehen, die einzelnen Skalenspannweiten für die Einzelkomponenten aber sehr viel schmaler sind als für den MCS und PCS und sich dann bei der 15-Prozent-Schwelle entsprechend Skalenspannweiten entwickeln bzw. Schwellenwerte ergeben, die so weit von den Einzelkomponenten entfernt sind, dass man unter Umständen, wenn man den MCS und PCS betrachtet, keinen Unterschied mehr erkennen kann, obwohl die Einzelkomponenten, die eingehen, sehr wohl einen großen Unterschied darstellen. Das war der Punkt, den uns Bjorner, der Entwickler des SF-36, an die Hand gegeben hat.

Frau Dr. Behring (amt. Vorsitzende): Ich kann dazu gar nichts sagen. – Frau Wieseler, bitte.

Frau Dr. Wieseler: Vielen Dank. – Das ist ein Papier von QualityMetric, das ist richtig. Das ist die Gruppe, die den SF-36 betreut. Mit dieser Gruppe haben wir in engem Kontakt gestanden, als wir den Skalenrange für den SF-36 abgeleitet haben. Das Problem, das Sie schildern, dass der auf Normstichproben genormt ist, haben Sie in jeder Auswertung, die Sie vom SF-36 machen, unabhängig davon, welches Responsekriterium Sie verwenden. Da geht man im Allgemeinen so damit um, dass man die aktuellste Normstichprobe verwendet. Das haben wir hier auch gemacht.

Mein Problem mit diesem Papier ist, dass da Responseschwellen vorgeschlagen werden, die in keiner Art und Weise unseren Ansprüchen entsprechen. Die Responseschwellen, die vorgeschlagen werden, sind allein verteilungsbasiert. Sie liegen im Wesentlichen unterhalb der Schwellen, die ich mit Messfehlern überhaupt noch entdecken kann. Das ist ein Problemkomplex in diesem Papier, der es für mich schwierig macht, mit den Responseschwellen, die hier vorgeschlagen werden, überhaupt zu arbeiten.

Zu dem Problem, das adressiert wurde, dass die Tatsache, dass ich auf einzelnen Skalen eine Response sehe, nicht dazu führt, dass ich auf der Gesamtskala eine Response sehe: Die Fallkonstellation, die da geschildert wird, ist, dass ich auf vier Subskalen eine Response sehe, dass in die Summenscores aber jeweils alle acht Subskalen eingehen. Die Annahme ist auch, dass auf den anderen Subskalen keine Veränderung vorliegt. Dann wundert es mich auch nicht, dass ich auf der Gesamtskala nichts sehe. Das ist ein Verhalten, das wir immer wieder haben, wenn die Summenskala zusätzlich Komponenten enthält, wo wir keine Änderung sehen. Das ist nichts, was für mich unser Vorgehen infrage stellt.

Wie gesagt, das Hauptproblem, das ich mit diesem Papier habe, ist, dass die vorgeschlagenen Responseschwellen verteilungsbasiert unterhalb der Detectible Changes liegen und das für mich infrage stellt, wie relevant dieser Einwurf ist.

Frau Dr. Behring (amt. Vorsitzende): Frau Huschens.

Frau Dr. Huschens (Janssen-Cilag): Vielen Dank, Frau Wieseler, das passt ganz gut dazu. Der Summenscore wird zwar aus allen acht Komponenten gebildet, aber wenn ich zum Beispiel den PCS bilde, werden die vier Komponenten, die die physikalische Funktion beschreiben, mit einer höheren Gewichtung eingehen. Das heißt, wenn ich da einen Unterschied habe und in den mentalen Eigenschaften weniger, dann sollte sich trotzdem beim PCS etwas darstellen, wenn ich in den physikalischen Eigenschaften etwas erkenne. – Danke.

Frau Dr. Behring (amt. Vorsitzende): Herr Hennig.

Herr Dr. Hennig (GSK): Zu dem validen Punkt von Ihnen, Frau Wieseler, dass Sie sich die Kriterien angeschaut haben und gesagt haben: Verteilungsbasiert ist vielleicht nicht die beste Methode. Es ist genau die Art der Diskussion, die wir uns wünschen, dass wir uns über die richtige Methode unterhalten und nicht über die Setzung eines Schwellenwertes. Es gibt – das haben Sie auch mit Ihrem Methodenpapier ausgeführt – eine ganze Reihe von Literatur zur Frage, ob es ankerbasiert oder verteilungsbasiert sein sollte. Ich glaube, dass diese Diskussion genau die richtige ist, die wir führen sollten, und nicht die Diskussion über einen Schwellenwert, der gesetzt wurde.

Frau Dr. Behring (amt. Vorsitzende): Möchten Sie replizieren, Frau Wieseler?

Frau Dr. Wieseler: Damit können wir vielleicht tatsächlich ein neues Thema einläuten, das schon mehrfach angesprochen wurde. Was ist das überhaupt für ein Wert, den wir vorgeschlagen haben, und ist er adäquat? Es ist schon mehrfach angeklungen und wird von verschiedenen Stellungnehmenden immer wieder betont, dass die MID von verschiedenen Faktoren abhängig ist und Sie deshalb abhängig von der Patientenpopulation unterschiedliche MIDs einsetzen möchten. Auch die Fachgesellschaften weisen darauf hin. Zum Beispiel beschreibt die Atemwegsliga, dass das von der Indikation abhängt, dass zum Beispiel für den SGRQ in der Lungenfibrose wahrscheinlich eine andere MID gilt als in der COPD, dass das abhängig ist von dem Schweregrad der Erkrankung, von dem Setting, das wiederum den Schweregrad abbildet. Ein Patient, der ambulant behandelt wird, hat wahrscheinlich eine andere MID als ein hospitalisierter, schwer erkrankter Patient oder jemand, der in der Reha ist. Auch die DDG weist darauf hin, dass die MID vom Schweregrad der Erkrankung abhängen kann. Das ist richtig. Das Problem ist nur, dass die MIDs, die uns aktuell vorliegen und mit denen bisher gearbeitet wurde, diese Variabilität nicht abbilden. Die Validierungsstudien, die für diese MIDs gemacht wurden, erfüllen die Ansprüche, die aktuell in der Literatur an die MIDs formuliert werden, nicht.

In dieser Situation kann man zwei Dinge tun. Man kann jede MID ablehnen, weil sie die aktuellen methodischen Anforderungen nicht erfüllt, und gar keine Responderanalysen mehr machen, oder man kann versuchen, einen Ausweg zu finden, der uns erlaubt, Responderanalysen zu machen, weil wir damit besser verstehen, was der Zusatznutzen eines Präparats ist.

Wir sind den zweiten Weg gegangen. Wir haben systematisch empirisch ermittelt, wo die MIDs im Moment liegen. Wir haben einen Wert vorgeschlagen, der am oberen Rand dieses empirisch ermittelten Felds liegt. Es ist richtig, das ist eine Setzung, aber das ist eine Setzung auf Basis der aktuell verfügbaren Empirie, Evidenz, basierend auf der Diskussion, die aktuell methodisch zu diesem Thema geführt wird, also eine Setzung auf Basis des aktuellen Stands der wissenschaftlichen Erkenntnis.

Was ich ein bisschen schwierig finde, ist, dass Sie auf der einen Seite beschreiben, die MIDs müssten die Variabilität abbilden, gleichzeitig aber auf MIDs zurückgreifen wollen, die das genau nicht tun. Die Alternativvorschläge, die es zu der 15-Prozent-Schwelle gibt, kann man aus meiner Sicht in drei Gruppen gliedern. Die eine Gruppe sagt: Lassen wir alles beim Alten. Die zweite Gruppe sagt: Wir möchten die alten, die wir bisher verwendet haben, verwenden, und für die Bereiche, wo wir keine MID haben, können wir die 15 Prozent nehmen. Die dritte Gruppe sagt: Lasst uns darüber reden, welche Lösung es gibt. Das Dritte findet statt, auch viele der hier Anwesenden sind beteiligt, zum Beispiel am SISAQOL-Projekt, wo diese Diskussionen geführt werden. Das wird noch einige Jahre in Anspruch nehmen. Das hilft uns heute konkret nicht weiter, wo wir alle zusammen jedes Jahr 80 bis 90 Bewertungen abwickeln müssen. In dieser Situation ist die Frage: Wie können wir weiter vorgehen? Da halten wir nach wie vor den Vorschlag der 15 Prozent, der keine MID darstellt, sondern einen empirisch gestützten Wert etwas oberhalb der MID, der aber immer noch eine kleine Änderung abbildet, für zielführend. – Danke schön.

Frau Dr. Behring (amt. Vorsitzende): Vielen Dank. Ich glaube, das war eine wichtige Erläuterung. – Herr Kardos.

Herr Dr. Kardos (Deutsche Atemwegsliga): Vielen Dank. – Ich bin Kliniker. Zu den statistischen Ausführungen kann ich nur am Rande Stellung nehmen. Ich unterstütze die Bemerkung von Frau Wieseler soeben, dass für verschiedene Krankheitsbilder selbst in dem schmalen Gebiet der Pneumologie verschiedene Schwellen gelten sollten. Dann ist aber meine Frage: Wieso sollte man für alle erdenklichen Skalen, für die die Metaanalyse durchgeführt wurde, die gleiche 15-Prozent-Responseschwelle nehmen? Ich meine, das ist nicht am oberen Rand zum Beispiel von St. George's Respiratory Questionnaire, was viel angibt. In anderen Untersuchungen geht das bis zu 7, bei Fremdzitaten und Beurteilungen. Der ursprüngliche Paul Jones hält sich bei 4 fest. Ich meine, es ist bei verschiedenen Populationen durchaus denkbar, verschiedene Schwellen anzusetzen. Aber wenn Sie die 15 Prozent zum Beispiel bei dem St. George's Respiratory Questionnaire ansetzen, dann haben Sie nicht eine einzige pneumologische Studie, die diese Responseschwelle erfüllt hätte. Bei uns in der Pneumologie sind der Inbegriff der Wirksamkeit die Biologika bei Asthma, die wirklich das Leben der Patienten, die dafür geeignet sind und die ansprechen, verändern. Das ist etwas ganz anderes als das, was zuvor war. Bei diesen Biologika sind Schwellenwerte in St. George's Respiratory Questionnaire – ich habe das leider in meiner Stellungnahme nicht geschrieben, aber jetzt habe ich die Literatur herausgesucht – für Benralizumab und Mepolizumab bei 7 bis 8, niemals bei 15. Bei 15 Prozent wären es bei SGRQ 15 Einheiten. Das ist nie erreicht worden. Muss ich, wenn diese 15 Prozent oder 15 Punkte auf SGRQ richtig sein sollten, das so interpretieren, dass Patienten, die mit Biologika behandelt sind, gar nicht davon profitieren, weil sie weit unter der Schwelle liegen? Das führt zu einem therapeutischen Nihilismus. Ich spreche gar nicht mehr über COPD, wo ohnehin in der Pharmakotherapie nur kleine Veränderungen möglich sind. Das ist sicherlich ein Problem.

Mein letzter Punkt wäre, dass die PROs, was wir hier mit MID oder Responseschwelle messen wollen, für die Ankermethode – unterer und oberer Anker – und nicht die Verteilungsmethode sind, die eine rein epidemiologische Bedeutung hat. Diese Responseschwelle ist aus statistischen Überlegungen entnommen worden, die noch viel weiter von den Patienten entfernt sind, die ohnehin Mittelwerte MCID oder MID sind. Aber wenn ich auf SGRQ zurückgreife, sind diese Werte zumindest am Patienten erhoben, und die

Patienten haben damit etwas zu tun. Die rein statistische Betrachtung und Simulation ist noch viel weiter vom Patienten entfernt. Das sind meine Bedenken. – Vielen Dank.

Frau Dr. Behring (amt. Vorsitzende): Herr Kardos, es gibt eine Rückfrage zu Ihren Ausführungen. Frau Wieseler, bitte.

Frau Dr. Wieseler: Vielen Dank. – Vielleicht zu Ihrem letzten Punkt, der Setzung der 15 Prozent auf Basis der Empirie. Diese Empirie waren alles Studien mit Patienten. Die Basis der Setzung ist das Patientenurteil. Das heißt, die Studien, die wir herangezogen haben, waren Studien zur Ermittlung der MID durch das Patientenurteil. Das ist systematisch auf Basis dieser Ergebnisse abgeleitet. Sie haben gefragt: Warum ist die Schwelle von 15 Prozent über alle Indikationen gleich, müsste sie nicht angepasst werden? Da wäre meine Frage: Wie soll man das festlegen? Gibt es das jetzt aktuell so, dass wir das einsetzen könnten? Das ist gegebenenfalls etwas, was auf den Punkt zielt: Man muss das alles einmal diskutieren. Aber das dauert, und wir brauchen jetzt eine Lösung.

Was Ihr Beispiel der Biologika bei Asthma angeht, da haben Sie von 7 bis 8 Punkten Unterschied gesprochen. Sind das Gruppenunterschiede? – Das sind Unterschiede zwischen Behandlungsgruppen.

(Herr Dr. Kardos (Deutsche Atemwegsliga): Ja!)

– Davon wäre ich auch ausgegangen. Vielen Dank für die Bestätigung. Wir meinen aber patientenindividuelle Unterschiede. Wenn ich einen Gruppenunterschied von 7 bis 8 Punkten sehe, wird es eine Verteilung der Patienten geben. Da wird es Patienten geben, die sehr viel höhere Änderungen haben. Das ist mit der Grund, warum wir uns gern die Empirie ansehen möchten. Wir würden gern die Anwendung des SGRQ ausgewertet sehen, wie viele Patienten diese Hürden nehmen und wie viele nicht, um genau solche Fragen beantworten zu können. – Vielen Dank.

Frau Dr. Behring (amt. Vorsitzende): Herr Worth, Sie sind dran.

Herr Prof. Dr. Worth (Deutsche Atemwegsliga): Ich möchte Peter Kardos noch etwas ergänzen in seiner Sorge zum Schwellenwert von 15 Prozent, gerade was den SGRQ betrifft, der bei unseren COPD-Patienten sehr häufig auch international untersucht und geprüft worden ist. Es hat sich zumindest für die bei Pharmastudien meist untersuchten Patienten, die Anfang 60 sind und relativ klar zwischen mittelschwerer und schwerer COPD liegen, gezeigt, dass ein Schwellenwert, eine MID von 4 ein eindeutig klinisch relevanter Effekt ist. Das heißt, das unterscheidet Patienten, ob sie zu Hause zurechtkommen müssen oder ob sie zum Beispiel einkaufen gehen können. Es ist für uns Kliniker eindeutig ein klinisch relevanter Punkt. Insofern ist der Wert von 15, der sehr weit davon weg liegt, selbst von unseren positiven Reha-Effekten, die bei 7 bis maximal 8 liegen, sehr weit weg, schwer zu verstehen, warum der 15-Prozent-Wert für alle Situationen benutzt wird. Ich stimme Frau Wieseler ausdrücklich zu, dass die Schwellenwerte durchaus von der Art des Kollektivs abhängen und dass diese jeweils mitbeschrieben werden müssen und vielleicht auch die Extreme in den Patientenkollektiven angegangen werden müssen, um zu sehen, wie da die MIDs liegen. Aber mit einem einheitlichen Wert von 15 Prozent alles zu beschreiben, da habe ich als Pneumologe Bauchschmerzen.

Frau Dr. Behring (amt. Vorsitzende): Ich würde gerne noch Herrn Hennig das Wort zum SGRQ geben und dann diesen Bereich abschließen. Herr Hennig, bitte.

Herr Dr. Hennig (GSK): Vielen Dank, Frau Behring. – Herr Kardos hat gerade angesprochen, dass der SGRQ von Professor Paul Jones entwickelt wurde; mit dem haben wir auch Kontakt aufgenommen. Ähnlich wie bei dem SF-36 haben wir uns an den Entwickler gewendet mit der Frage, ob in dieser spezifischen Konstellation ein Schwellenwert von 15 Prozent relevant ist. Wir haben das auch eingereicht. Das Kernargument vom Entwickler des Fragebogens war, dass es eine Reihe von Studien gibt, die die Bedeutung auch für die Patienten von der Schwelle von 4 ablehnen. Wir reden von einem Unterschied von 4 zu 15. Des Weiteren hat sich

Professor Jones mit einem relativ aktuellen Papier von Devji und Kollegen, erschienen 2020, auseinandergesetzt und hat die Kriterien zugrunde gelegt und kommt zu der gleichen Einschätzung, dass die 15 wirklich keinen Sinn macht, auch aus Entwicklerperspektive. Ich wollte das zu Protokoll geben, weil Sie ganz am Anfang die Stellungnehmer vorgelesen haben. Wir hatten mit unserer Stellungnahme auch eine Stellungnahme von Professor Jones eingereicht, die ich jetzt ganz kurz zusammenzufassen versucht habe.

Frau Dr. Behring (amt. Vorsitzende): Ich bin dankbar, dass nicht jeder seine Stellungnahme in Gänze vorliest. – Herr Müller-Wieland, bitte.

Herr Dr. Müller-Wieland (DDG): Ich wollte die breitere Perspektive aus klinischer Sicht einnehmen. Erst einmal ganz herzlichen Dank, Frau Wieseler, für die ausführliche Darlegung. Jetzt können wir besser nachvollziehen, wie Sie zu dem Vorschlag der 15 Prozent gekommen sind. Zu dem Zeitpunkt, als wir bei uns im Kreise den Vorschlag diskutiert haben, war der Punkt aus unserer klinischen Sicht, ich glaube, auch getragen durch das Argument von Frau Müller, ab wann oder wie ist eigentlich eine klinische Relevanz festzulegen, ob die Lösung für das Thema rein methodisch-strategisch sinnvoll ist. Es ist aus klinischer Sicht für uns nicht ganz nachvollziehbar, zu sagen, man legt in dieser Situation einen Einheitswert als Relevanzschwelle für etwas vor – ich glaube, darüber gibt es auch Einvernehmen –, was keine Einheit hat. Dadurch kommt die Diskussion der Verzerrung. Selbstverständlich können wir gut nachvollziehen, dass die verschiedenen PROs – wir müssen eine Handhabe haben, und es werden sich viele neue entwickeln – unterschiedlich evaluiert sind. In unserem Gebiet, von Diabetes über Herzinsuffizienz, kann man relativ klar feststellen, dass in der Literatur manche Fragebögen sind, die sehr gut evaluiert sind, auch in Bezug auf Schwellenwerte, bezogen auf ein klinisch relevantes Ausmaß. Ein paar Beispiele haben wir eben schon gehört. Andere sind noch gar nicht oder schlecht evaluiert. Insofern denken wir, dass man sich seitens des IQWiG methodisch festlegen könnte, dass man sich auf klinisch zumindest akzeptabel evaluierte Fragebögen bezieht. Dann kann man zu dem Punkt kommen – Frau Wieseler, das kann ich auch nachvollziehen –, zu sagen: Je nachdem reicht das, was als klinisch relevantes Ausmaß seitens des Fragebogens sich herauskristallisiert, nicht für das Verfahren der Zusatznutzenbewertung. Das ist letztlich der Hintergrund, dass wir auf die 15 Prozent kommen. Auch das kann ich nachvollziehen. Wenn man das dann lösungsorientiert vorschlägt – das ist der Punkt –, muss ich die Differenz oder der Abstand zu den bereits bestehenden Daten nicht einheitlich formulieren, dann müssen Sie es auf die obere Grenze der Verteilung beziehen. Das heißt, dann kann man einen Prozentsatz x nehmen, aber er muss sich auf die Evaluierung beziehen bzw. auf den individuellen Evaluierungsbogen.

Auch wenn es kein konkreter Lösungsvorschlag ist: Das sind unsere Bedenken, und die mehren sich. Wir haben schon Über- oder Unterbewertungen. Aber natürlich braucht man eine reale Handhabe bei der großen Entwicklung der verschiedenen Fragebögen bei dem Thema PROs.

Unser Punkt wäre: keine Einheitsschwelle für etwas, was nicht einheitlich ist, sondern sich auf eine Differenzschwelle zu beziehen, die sich in der Differenz auf die Metrik und die klinische Evaluierung der einzelnen Fragebögen und damit deren individueller Skalierung bezieht. Bei allem anderen verlieren wir die Individualität, und dann kommen wir zu den Problemen, die hier besprochen worden sind. – Vielen Dank.

Frau Dr. Behring (amt. Vorsitzende): Frau Wieseler, möchten Sie noch einmal dazu Position beziehen?

Frau Dr. Wieseler: Ja, vielen Dank. Ich hätte sogar eher eine Nachfrage. – Vielen Dank, Herr Müller-Wieland. Sie schlagen im Grunde genommen eine individualisierte Schwelle für die Fragebögen vor und haben von dem oberen Rand einer Verteilung gesprochen. Mir ist nicht ganz klar geworden, welche Verteilung Sie meinen. Meinen Sie die Verteilung der verfügbaren MIDs für ein Instrument? Denn wir wissen aus der Arbeit von Devji, dass es für viele Instrumente viele MIDs gibt. Oder meinen Sie die Verteilung der Werte bei Patienten? Vielleicht könnten Sie das für mich noch klären.

Herr Dr. Müller-Wieland (DDG): Frau Wieseler, Sie sagen: Wir schlagen eine Relevanzschwelle vor. Jetzt ist die Frage: Wo und wie zieht man diese Relevanzschwelle? Mit allem Vorbehalt der unterschiedlichen Fragebögen: Der entscheidende Punkt ist, dass Sie sagen: In aller Regel entspricht die Relevanzschwelle der einzelnen PROs nicht dem Anspruch für das Zusatznutzenbewertungsverfahren. Das heißt, ich kann nicht aus der hohlen Hand für alle Fragebögen sagen: Beziehen wir uns auf die Verteilung. Aber wenn es evaluiert ist in Bezug auf ein klinisches Ausmaß, wie wir eben hörten, ist die Zahl 4 ein Schwellenwert, der klinisch evaluiert ist; machen dann eine prozentuale Verschiebung nach oben, aber bezogen auf den evaluierten oder angenommenen Schwellenwert der klinischen Relevanz der einzelnen Fragebögen. Wenn es bei dem einen 25 ist, dann sagen Sie: 25 reicht mir aber nicht, ich wähle 30. Das ist nachvollziehbar, bezogen auf die Relevanzschwelle des individualisierten Fragebogens. Dadurch bekommen Sie eine Individualisierung und nicht eine Einheitsschwelle. Ich kann es nachvollziehen, aber ob das die beste Lösung ist, bezweifle ich.

Frau Dr. Wieseler: Das Problem, das da besteht, ist die Validität der jetzt vorliegenden MIDs. Das ist etwas, was massiv infrage steht.

Herr Dr. Müller-Wieland (DDG): Die sind sehr unterschiedlich zwischen den einzelnen Fragebögen. Und sehr konkret: Es gab zuletzt Nutzenbewertungsverfahren, wo die verwendeten PROs sehr gut evaluiert sind. Vielleicht ist da überhaupt kein Bedarf für eine zusätzliche Annahme. Deswegen der Punkt, zu sagen: Vielleicht sollte man zunächst einmal indikations- und populationsbezogen, zumindest indikationsbezogen die vorhandenen PROs in Bezug auf die Qualität der Evaluierung aus Sicht des IQWiG gruppieren. Dann kann man sich gegebenenfalls strategisch festlegen und nachvollziehen, auf welche Sie sich beziehen. Dann ist für die pharmazeutischen Unternehmer transparent, welche in den verschiedenen Untersuchungen verwendet werden könnten/sollten. Wenn sich neue PROs entwickeln, ist klar, was der Transparenzkatalog für eine klinische Evaluierung ist. Aber ich bleibe dabei: Für die Relevanzschwelle einer Nutzenbewertung über alle PROs von chronisch obstruktiver Lungenerkrankung, Niereninsuffizienz, Herzinsuffizienz, Diabetes einen einheitlichen Prozentsatz zu wählen, eine Relevanzschwelle, damit werden wir dem Problem nicht gerecht. Das war auch der Punkt, das war die Ausgangslage.

Frau Dr. Behring (amt. Vorsitzende): Frau Wieseler, ein letztes Statement dazu, oder würden Sie das so stehen lassen?

Frau Dr. Wieseler: Wir drehen uns dahin gehend im Kreis, als der Ausgangspunkt für diesen Vorschlag, Herr Müller-Wieland, ist, dass wir eine hinreichend valide Basis haben. Das steht infrage. Das ist unser Problem.

Frau Dr. Behring (amt. Vorsitzende): Herr Kiencke, bitte.

Herr Dr. Kiencke (Novo Nordisk): Ich habe eine Anmerkung zu Frau Wieseler. Wir sehen schon ein, dass es lösungsorientierte Ansätze geben muss. Die MIDs sind nicht in Gänze gut entwickelt, oder es ist historisch fragwürdig, ob das modernen Ansprüchen entspricht. Sie hatten vorhin ausgeführt, dass Sie systematische Recherche über Arbeiten gemacht haben, die sich mit dem Thema MID/MICD beschäftigen, und da haben Sie die 15 Prozent abgeleitet. Ich habe alle diese Arbeiten studiert und habe einen anderen Eindruck gewonnen. Das sind ausgewählte Indikationsgebiete. Das ist sehr eingeschränkt. Ich komme nicht zu dem Schluss, dass Sie die 15 Prozent regelhaft abgeleitet haben. Dieses Thema sind Sie eben angegangen. Es wäre hilfreich, wenn Sie systematisch darstellen könnten, wie Sie zu den 15 Prozent kommen. – Das als konstruktiver Ansatz, weil ich glaube, die Setzung – Sie sagten, es sei eine Setzung – ist vielleicht doch nicht so systematisch abgeleitet. Meiner Einschätzung nach war es auch keine systematische Literaturrecherche, die für die Nutzenbewertung notwendig wäre, sondern eher eine fokussierte Recherche.

Frau Dr. Behring (amt. Vorsitzende): Herr Kiencke, eigentlich hatte ich den Eindruck, dass Frau Wieseler das schon deutlich erläutert hatte. Frau Wieseler, möchten Sie noch einmal dazu Stellung nehmen, es kurz wiederholen?

Frau Dr. Wieseler: Ich kann es noch einmal kurz darstellen. Es ist richtig, es ist eine fokussierte Recherche nach systematischen Reviews mit Studien zur Ermittlung von MIDs. Diese systematischen Reviews haben wir herangezogen, um die einzelnen Studien zur Ermittlung der MIDs zu prüfen und darzustellen, in welchem Bereich die MIDs lagen. Das Ganze ist im Detail in der Würdigung der Stellungnahmen zum Methodenpapier dargestellt – wenn noch Bedarf ist, sich das anzuschauen.

Was die Studien angeht, haben wir die Studien ausgewählt, die zumindest annäherungsweise den heutigen Ansprüchen entsprechen. Es mussten also Longitudinalstudien sein, in denen ankerbasierte MIDs ermittelt wurden. Der Cut-off musste eine kleine Änderung zeigen. Das sind die Anforderungen, die aus der McMaster-Gruppe für die MID-Studien formuliert werden. Auf Basis dieser systematischen Arbeiten haben wir dann – auch das finden Sie in dem Dokument – dargestellt, wie der Range der empirisch ermittelten MIDs aussieht. Hier ist mehrfach angeklungen, wir hätten irgendeinen Mittelwert gewählt. Nein, wir haben auf Basis dieser Empirie – das ist aktuell die bestverfügbare Evidenz dazu – eine empirisch gestützte Setzung vorgenommen, weil aus unserer Sicht im Moment kein anderer Vorschlag zu diesem Zeitpunkt einen machbaren Weg für die Gesamtheit der Nutzenbewertungsverfahren darstellt. Es ist unbenommen, dass, wenn die Diskussion weitergeht – das tut sie; ich habe zum Beispiel die SISAQOL-Gruppe genannt –, dieses Verfahren angepasst wird. Aber nach dem, was wir aktuell haben, haben wir das für den besten Weg gehalten.

Frau Dr. Behring (amt. Vorsitzende): Vielen Dank noch einmal. – Frau Teupen, bitte, Ihre Frage.

Frau Teupen: Herr Professor Wörmann ist leider nicht da, ich weiß nicht, ob jemand anderes die Frage beantworten kann. Er bezieht sich in der Stellungnahme im Rahmen der AWMF in der Kommission Nutzenbewertung Arzneimittel darauf, dass sie es begrüßen würden, dass es Veränderungen/Verbesserungen für die MIDs gibt. Er sagt aber auch, dass eventuell möglich sein sollte, für kleine Patientengruppen, ... [akustisch nicht zu verstehen], aber auch Orphans, auch Superorphans, größere Spannweiten akzeptiert werden müssten. – Auch wenn wir grundsätzlich ein Problem haben mit Validierung, kann man das nicht übertragen auf die spezifische Erkrankung. Ich weiß nicht, wie man das für die seltenen Erkrankungen sieht, insbesondere Orphans. Wie sehen das die Teilnehmer? Er hat vorgeschlagen, breitere Spannweiten zu ermöglichen.

Frau Dr. Behring (amt. Vorsitzende): Sieht sich jemand berufen, zu den Orphans etwas zu sagen? – Herr Sauerbruch ist von der Inneren Medizin. Möchten Sie sich gleichzeitig zu Orphans äußern? Wahrscheinlich ist das eher eine andere Wortmeldung.

Herr Prof. Dr. Sauerbruch (DGIM): Zu den Orphans kann ich jetzt nichts sagen. Das zeigt aber insgesamt, wie schwierig das Feld ist und wie schwierig es ist, zu erfassen, ob es dem Patienten gut oder schlecht geht. Was ist das Ziel? Wir wollen wissen, ob es unter einer Behandlung dem Patienten besser oder schlechter geht, und wir wollen das irgendwie erfassen. Sie sehen natürlich auch bei einem Orphan Disease: Wenn sich der Arzt mit dem Patienten unterhält, können sie das gemeinsam erfassen.

Was ich aus dieser endlosen und manchmal schwer nachzuvollziehenden statistischen Diskussion gelernt habe, ist, dass wir a) nicht immer die Instrumente haben, um in den verschiedenen Lebensqualitätsscores zu sehen, ob es dem Patienten schlechter oder besser geht, und dass wir b) in der statistischen Auswertung des Ganzen große Schwierigkeiten haben. Frau Wieseler hat es am Ende schön zusammengefasst: Warum lässt man nicht den 15-Prozent-Schwellenwert der Spannweite einfach mitlaufen, ohne ihn als Entscheidungskriterium heranzuziehen, jetzt am Anfang? Man sieht dann, wie sich das

entwickelt und ob man damit die Realität besser erfassen kann, um das später in die Entscheidung hineinzunehmen, einfach parallel laufen. Nach dieser langen Diskussion schien mir das alles sehr offen zu sein. – Vielen Dank.

Frau Dr. Behring (amt. Vorsitzende): Vielen Dank, Herr Sauerbruch. Wir haben es seit letztem Dezember parallel laufen lassen. Es wurde uns nicht immer beides dargestellt. Danke für Ihren Appell. Das hat uns unterstützt. – Frau Müller, Sie können gerne dazu etwas sagen.

Frau Dr. Müller: Sie haben gerade gesagt, Frau Behring, was ich sagen wollte.

Frau Dr. Behring (amt. Vorsitzende): Wir hatten versucht, die Wahrheit zu finden, Herr Sauerbruch. Es ist nicht so leicht, wie Sie gerade feststellten. – Herr Rasch, Sie haben sich gemeldet.

Herr Dr. Rasch (vfa): Zu den Ausführungen von Frau Wieseler, dass die 15 Prozent systematisch abgeleitet wurden: In der Dokumentation zu dem Stellungnahmeverfahren ist die Recherche beigelegt. Es ist eine Bandbreite. Für bestimmte Indikationen wurde irgendwie in der Mitte ein Mittelwert gesetzt. Okay, das ist eine sehr pragmatische Lösung. Das kann man nachvollziehen. Aber die Einheitsschwelle ist gerade nicht der wissenschaftliche Konsens. Ich will ganz klar dem widersprechen, dass alle MIDs international massiv in der Kritik stehen. So ist es nicht. Alle HTA-Organisationen und auch Zulassungsbehörden nutzen weiterhin etablierte MIDs. Es ist nicht so, dass diese MIDs rigoros überall abgelehnt werden. Noch viel wichtiger ist: Niemand nutzt aktuell eine Einheitsschwelle. Es wäre ein ganz klarer Alleingang, wenn man jetzt eine Einheitsschwelle nehmen würde. Ich habe die Stellungnahmen der Fachgesellschaften genau gelesen. Da wird es genauso gesehen. Eine Einheitsschwelle wird nicht als sinnvoll angesehen. Es ist sehr bedauerlich, wenn bei internationalen Instrumenten wie SF-36 oder QLQ-C30 die MIDs abgelehnt werden, ohne dass sie überhaupt geprüft werden. Nach dem neuen Methodenpapier soll keine Prüfung mehr stattfinden. Es wird immer wieder gesagt, es gibt keine Prüfkriterien. Das ist irreführend. Wir sind nicht im luftleeren Raum. Wir haben seit Jahren, seit Jahrzehnten Prüfkriterien, gängige Qualitätskriterien. Wir haben zwar keinen Goldstandard; das ist richtig – von daher kommt das IQWiG –, aber es sind sinnvolle Prüfkriterien, die auch vom IQWiG jahrelang in die Diskussion eingebracht wurden. Die müssen konzentriert und weiterentwickelt werden. Das sind genau die Stichpunkte, die heute erwähnt wurden: Längsschnittstudien, ankerbasierte Verfahren, Korrelationen. Da streben wir eine Diskussion und letztlich einen Konsens an, womit man dem internationalen Standard letztlich Rechnung tragen würde, aber keinen Alleingang, der von allen Seiten als nicht tragbar angesehen wird. – Danke.

Frau Dr. Behring (amt. Vorsitzende): Vielen Dank, Herr Rasch. Das geht in die Richtung, die Herr Hennig sagte. Sie sind alle gewillt, neue Validierungsstudien von den MIDs zu machen. Wir haben bisher noch nicht so richtig gesehen, dass tatsächlich positive Validierungsstudien hinterhergeschoben werden. – Frau Wieseler.

Frau Dr. Wieseler: Herr Rasch, es ist nicht so, dass wir nicht wüssten, welche Kriterien herangezogen werden. Auf der anderen Seite ist das gar nicht das Problem. Wenn Sie sagen, dass wir die Responsekriterien vom SF-36 nicht in jedem Verfahren neu prüfen, dann muss ich Ihnen sagen: Ein verteilungsbasiertes Verfahren auf einer Normstichprobe, um ein Responsekriterium festzulegen, ist definitiv nicht State of the Art. Das brauchen wir nicht jedes Mal wieder zu prüfen.

Die Kriterien sind seit langer Zeit in der Diskussion. Sie sind im letzten Jahr erstmals zusammengefasst worden, auch von der McMaster-Gruppe, die sich seit 30 Jahren damit beschäftigt. Die haben auch eine Empirie gemacht. Sie haben eine Vollerhebung gemacht und über Jahre systematisch alle Studien zur MID-Validierung identifiziert und haben untersucht, ob diese Studien die Qualitätskriterien treffen. Das Ergebnis dieser empirischen Untersuchung ist sehr ernüchternd. Die Studien untersuchen in der Regel nicht die Kriterien, die notwendig sind, um die Qualität der MIDs festzulegen. Das ist etwas, was sich mit unserer Erfahrung

deckt. Wir sind diesen Schritt nicht aus purem Übermut gegangen. Vielmehr haben wir in den Verfahren festgestellt, dass die MID-Validierungsstudien den modernen, aktuellen Anforderungen nicht genügen. Die empirische Arbeit von Carrasco-Labra in „Journal of Clinical Epidemiology“ 2021 unterstützt das. Die beschreibt, dass die aktuell vorliegenden MID-Validierungsstudien nicht ermöglichen, die Validität dieser MIDs zu beurteilen.

In dieser Situation gehen wir einen Schritt, der es trotzdem ermöglicht, mit diesen Responderanalysen zu arbeiten. Das ist die Motivation dazu. Da hilft es mir nicht, zu sagen: Es gibt Qualitätskriterien. Dann schaue ich in die Validierungsstudien und sehe: Ich habe die Information nicht. Es gibt keine Angaben zur Korrelation mit dem Anker. Die MIDs sind nicht präzise. Wir haben nach wie vor Publikationen, wo MIDs aus zweimal 20 Patienten abgeleitet werden. Das ist etwas, was wir nicht ignorieren können.

Frau Dr. Behring (amt. Vorsitzende): Herr Leverkus, bitte.

Herr Leverkus (Pfizer): Ich wollte darauf hinweisen, es gibt immer wissenschaftlichen Fortschritt, die Kriterien werden anders. Das heißt, wenn wir in Zukunft Validierungsstudien machen, wird man die nach anderen Kriterien machen. Aber das heißt nicht, dass man das, was da war, überhaupt nicht mehr verwenden kann. Ich kann mich noch daran erinnern, dass wir vor einigen Jahren über etablierte MIDs gesprochen haben. Die MIDs, die wir bis jetzt in Verfahren seit 2011 eingesetzt haben, sind in der Regel als validiert anzusehen. Es ist ein Einzelfall, wenn Sie im Prinzip sagen: Das akzeptieren wir nicht mehr. Es wäre etwas anderes, wenn auch die Zulassungsbehörden FDA und EMA sagen würden: Die MIDs von dieser Skala, das geht nicht mehr durch. Nachdem man die Arbeit von McMaster‘s gelesen hat, kann man das nachvollziehen. McMaster‘s ist im Prinzip eine Gruppe, die nicht den breiten Konsens hat, dass alle mitmachen. ... [Tonausfall] Kollegen, die diese Fragebögen entwickeln, sagen: Jetzt haben wir gesehen, dass dieser MID nicht valide ist, jetzt machen wir eine neue Validierungsstudie. Aber das passiert auch nicht. Ich kann mir gut vorstellen, dass, wenn wir nichts haben, die 15 Prozent ein Punkt ist, mit dem man vorwärtsgeht, damit man weiterkommt, aber dass man die alten, etablierten MIDs, die seit zehn Jahren in der Nutzenbewertung eingesetzt werden, gar nicht mehr anerkennt, das erscheint mir ein bisschen schwierig. Denn von den Zulassungsbehörden sehen wir das im Prinzip nicht, und von den Entwicklern haben wir das auch nicht gesehen. – Danke schön.

Frau Dr. Behring (amt. Vorsitzende): Bitte, Herr Hennig.

Herr Dr. Hennig (GSK): Ich wollte auf die laufende wissenschaftliche Diskussion hinweisen. Frau Wieseler, Sie hatten zu Recht auf das Papier von Devji hingewiesen. Das ist sicherlich ein wichtiges Papier, das wir uns alle sorgfältig angeschaut haben, über das man diskutieren kann. Der wissenschaftliche Dialog läuft. Es gibt durchaus andere Stimmen, die sich zu bestimmten Kriterien anders äußern. Da geht es konkret um die Höhe der Korrelation, die man anlegen sollte. Der wissenschaftliche Dialog, der ganz wichtig ist, läuft, ist aber aus meiner persönlichen Sicht noch nicht abgeschlossen.

Frau Dr. Behring (amt. Vorsitzende): Danke. – Herr Dintsios.

Herr Dr. Dintsios (Bayer): Mich hat gewundert, was man mit dem Methodenpaper 6.0 aufgebracht hat. Es gibt in Deutschland das EbM-Netzwerk. Das IQWiG hatte die „IQWiG im Dialog“-Veranstaltung. Es gibt die Cochrane Workshops. Man hätte ein bisschen vorarbeiten können, zusammen mit der anderen Seite. Es ist nicht so, dass die Industrie allergisch auf das IQWiG reagiert, vice versa. Man kann den wissenschaftlichen Austausch auch suchen.

Das Zweite – es geht auch um die Frage von Frau Teupen –: Bei Orphan Drugs werden die Validierungsstudien, sofern sie gemacht werden, zweimal 20 Personen haben. Wenn die Zielkollektive klein sind, werden die Validierungsstudien noch kleiner sein. Ob sie belastbar, robust sind, stellt sich als Frage.

Noch einmal zu den 15 Prozent. Es gab auch andere Vorschläge, nämlich ausgehend von dem, was da war. Wenn das IQWiG mit einem Sicherheitsabstand arbeiten möchte – das ist das

Petitum, es ist keine MID, richtig, von Frau Wieseler noch einmal dargelegt –, das ist eine Art Sicherheitsabstand, eine Art extern gesetzte Schwelle. Wieso dann nicht innerhalb der Variation, die es gab, einen Aufschlag machen? Bei dem Beispiel, das wir vorhin gehört haben mit den 4 Prozent auf einer 100er-Sakla, ist 15 mehr als das Dreifache. Das darf man auch nicht unterschlagen.

Frau Dr. Behring (amt. Vorsitzende): Danke, Herr Dintsios. – Frau Müller.

Frau Dr. Müller: Ich habe immer noch nicht verstanden, dass gesagt wird, man soll, wie Sie, Herr Dintsios, das wieder vorgeschlagen haben, einen Sicherheitsaufschlag auf die alten MIDs machen, wo eben zu hören war, dass die Qualität der Validierung so unterschiedlich wäre. Ich habe nicht verstanden, was das für einen Benefit bringen soll. Man hat einen Sicherheitsaufschlag, aber bei den EORTC-Bögen – das wurde auch in den Stellungnahmen angesprochen – liegen die 10 Punkte bei den meisten ohnehin ungefähr bei 15 Prozent. Da hätte man einen Aufschlag, der dazu führen würde, dass man dort eine Verschärfung hat, gerade bei den etablierten, noch relativ gut validierten Instrumenten, auch wenn da noch gearbeitet wird. Bei anderen hätte man den gleichen Aufschlag und hat eine Validierungsstudie, von der alle gesagt haben, dass es problematisch ist, mit der man nicht viel anfangen kann. Was mir ein bisschen fehlt, ist ein Alternativvorschlag von Ihrer Seite. Sie sagen, man müsse sich zusammensetzen. Sie sagen aber auch, es gibt bisher keine Versuche in relevantem Ausmaß, noch nicht ausdiskutierte, aber von der Richtung her andere, validere Kriterien MIDs zu ermitteln. Die Frage ist: Wann passiert das? Passiert das, wenn weiterhin die alten MIDs akzeptiert werden, oder passiert das erst, wenn es passieren muss? Ich sage es etwas salopp. Dass daran gearbeitet wird, ist richtig. Der Diskussionsprozess ist noch im Gange und ist noch nicht abgeschlossen. Ich hoffe, er ist nie abgeschlossen. Das wäre bei einer wissenschaftlichen Diskussion äußerst bedenklich. Er ist also noch im Gange, aber eine Richtung und bestimmte Kriterien – die haben Sie alle genannt – sind eigentlich schon klar: dass man zum Beispiel keine verteilungsbasierten Verfahren anwendet. Trotzdem passiert wenig. Orphans nehme ich einmal aus; da ist es schwierig. Da gibt es aber eine Zustimmung für eine allgemeine Relevanzschwelle von 15 Prozent. Das ist, glaube ich, nicht so problematisch als Ersatz dafür. Aber was ist wirklich Ihre Alternative? Weiterhin die alten zu verwenden, weiterhin die alten mit einem Standardaufschlag, der für manche sogar noch härter wäre als die 15 Prozent, für andere sehr viel leichter zu reißen wäre, aber möglicherweise auf der Grundlage einer unzureichenden Validierung, in Absprache mit anderen HTAs? Das ist alles sehr verständlich, aber glauben Sie, dass da in absehbarer Zeit – ich frage nur – etwas dabei herauskommt und man sich einigen und damit arbeiten kann? Halten Sie das für realistisch?

Frau Dr. Behring (amt. Vorsitzende): Herr Leverkus, Sie wagen sich vor.

Herr Leverkus (Pfizer): Ich wage mich in das verminte Terrain. Was ich schade finde, ist, dass man die alten, etablierten gar nicht mehr anerkennt. Es ist nicht alles schwarz oder weiß. Es wird sicherlich einige geben, wo man sagt: Komisch, das sind nur zehn Patienten. Es gibt methodische Sachen. Ich denke an EORTC oder andere Skalen; das sind ganze Forschergruppen, die sich damit beschäftigt haben. Wenn sich die Kritik von McMaster‘s weiter durchsetzen würde, die Diskussion weitergeht, man zu einem Konsens der Methodenentwickler im Bereich der Skalen kommt, dann würden solche Zentren oder die EORTC-Gruppe sicherlich Studien auflegen. Wenn das Verständnis ist: Das wird diskutiert, und ob das alles richtig ist, wissen wir im Prinzip auch nicht!, passiert nichts, dann werden solche Leute ihre Ressourcen anders einsetzen. Ich habe schon Vertrauen in die Skalenentwickler, dass die, wenn es wirklich Probleme gibt, versuchen, die MID zu definieren. Das ist nicht nur in der Nutzenbewertung eine relevante Sache, sondern sicherlich auch in der Klinik.

Frau Dr. Behring (amt. Vorsitzende): Danke, Herr Leverkus. – Wir ringen tatsächlich um eine Lösung. Diese Plattform hat schon einiges gebracht und Probleme aufgeworfen. – Herr

Dintsios, die letzte Wortmeldung geht an Sie. Sie dürfen noch einmal zum Sicherheitsabstand Stellung nehmen.

Herr Dr. Dintsios (Bayer): Ich kann natürlich die Sachen nachvollziehen, Frau Müller. Aber was nicht wissenschaftstheoretisch der Fall sein kann, ist, ein Problem durch etwas lösen zu wollen, was andere Friktionen mit sich bringt. Das ist nicht zielführend, rein effizienzmäßig bringt es uns nicht weiter. Es würde zu mehr Problemen führen.

Frau Behring, dass Validierungsstudien nachgeliefert wurden: Sie wissen, dass Sie keine Validierungsstudien im Rahmen von Pivotalstudien sehen wollen. Das wissen wir genauso. Wir können in Pivotalstudien nicht Amendments in letzter Sekunde ändern. Das Ganze ist letzten Sommer besprochen worden. Dahin gehend ist es normal, dass wir nicht plötzlich mit solchen Validierungsstudien aufwarten. Wo sollen die herkommen? Wir haben eine Evidenz, die wir bei Ihnen einreichen, das sind pivotale Studie hauptsächlich. Da können wir nicht etwas hineinpflanzen. Herr Leverkus hat es sehr gut gesagt: Es gibt auf den internationalen Foren wissenschaftliche Entwickler, die akademisch verortet sind. Die müssen wir auch einmal ... [akustisch nicht zu verstehen]. Ich kann keinen EQ-5D oder QLQ-C30 nachvalidieren. Das ist deren Problem. Das ist deren Plattform. Da sind 40 Leute aus zwölf europäischen Ländern involviert. Wie wollen wir von außen sozusagen als Eingabe das lösen?

Frau Dr. Behring (amt. Vorsitzende): Es ist tatsächlich ein Dilemma. – Herr Rasch, bitte.

Herr Dr. Rasch (vfa): Ein letztes Wort von mir. Wir haben in allen Stellungnahmen gesehen, dass eine Einheitsschwelle nicht definiert wird. In der laufenden wissenschaftlichen Diskussion, die erwähnt wurde, war nicht die Lösung, stattdessen eine Einheitsschwelle zu nehmen. Die Lösung war: Wir müssen uns dieser Diskussion kritisch stellen und die Kriterien definieren. Man muss die Validierungsstudien prüfen. Natürlich wird es auch Validierungsstudien geben, die besser oder schlechter sind. Aber es ist nicht so, dass wir gar kein Instrumentarium haben. Auf der Erkenntnis, dass weltweit niemand eine Einheitsschwelle vorschlägt und dass im Stellungnahmeverfahren, auch von der AWMF ganz klar gesagt wurde, eine Einheitsschwelle ist nicht die Lösung, sehe ich persönlich keine Grundlage, auf diesen Erkenntnissen den neuen Standard für die Nutzenbewertung zu setzen.

Frau Dr. Behring (amt. Vorsitzende): Vielen Dank. – Vielen Dank für diese rege Diskussion. Gibt es von einem der Stellungnehmer etwas, was noch nicht angesprochen worden ist? – ich habe das Gefühl, dass vieles mehrfach angesprochen, aber unterschiedlich beleuchtet worden ist. Zusammenfassend lässt sich sagen, dass der Großteil der Stellungnehmer gegen die Einheitsschwelle ist, und alle plädieren für die verwendeten etablierten MIDs. Ich denke, dass das IQWiG genug Zeit hatte, das zu erläutern, und das bei dem einen oder anderen zu mehr Verständnis geführt hat, warum und wie hier vorgegangen worden ist.

Alles, was gesagt worden ist, werden wir weiter in unsere Diskussion im G-BA hineinnehmen, um zu entscheiden, wie die Modulvorlagen aussehen werden. Das wird sicherlich noch eine spannende Diskussion.

Vielen Dank für Ihre rege Teilnahme und auch für Ihre aufwendigen Stellungnahmen. Wir haben uns darüber gefreut – auch wenn es uns viel Arbeit gekostet hat –, auch über die Aufarbeitungen, die Sie gemacht haben. Ich möchte Sie motivieren, noch eine Weile beide Schwellenwerte parallel darzustellen, sowohl die 15 als auch die alte MID, solange wir die Modulvorlagen nicht geändert haben und dazu keine Entscheidung getroffen haben. Für uns im G-BA ist es sehr wichtig, diese Daten zu sehen und ein Gefühl dafür zu bekommen, auch wenn nichts statistisch auswertbar ist.

Ich wünsche Ihnen noch einen schönen Nachmittag. Danke für Ihre Disziplin und für das Zuhören. Bis dann! Wiederschauen!

Die Anhörung ist beendet.

Schluss der Anhörung: 11:53 Uhr

C. Anhang der Zusammenfassenden Dokumentation

Bekanntmachung des Beschlusses im Bundesanzeiger

1. Unterlagen des Stellungnahmeverfahrens

1.1 Schriftliches Stellungnahmeverfahren

Bundesanzeiger
Herausgegeben vom Bundesministerium der Justiz und für Verbraucherschutz
www.bundesanzeiger.de

Bekanntmachung
Veröffentlicht am Freitag, 25. Juni 2021
BAnz AT 25.06.2021 B5

Bundesministerium für Gesundheit

Bekanntmachung des Gemeinsamen Bundesausschusses gemäß § 91 des Fünften Buches Sozialgesetzbuch (SGB V)

Vom 17. Juni 2021

Der Gemeinsame Bundesausschuss (G-BA) hat am 17. Juni 2021 beschlossen, ein Stellungnahmeverfahren zur Änderung der Verfahrensordnung nach § 35a Absatz 3 Satz 2 in Verbindung mit § 92 Absatz 3a in Verbindung mit § 91 Absatz 4 Nummer 1 SGB V einzuleiten:

5. Kapitel der Verfahrensordnung: Änderung der Modulvorlage in der Anlage II:

Anpassung der Anlage II.6 (Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen) zur Konkretisierung der Ergebnisdarstellung von patientenberichteten Endpunkten zur Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen

Der G-BA hat gemäß § 91 Absatz 4 Satz 1 Nummer 1 SGB V eine Verfahrensordnung zu beschließen, in der er insbesondere methodische Anforderungen an die wissenschaftliche sektorenübergreifende Bewertung des Nutzens, der Notwendigkeit und der Wirtschaftlichkeit von Maßnahmen als Grundlage für Beschlüsse sowie die Anforderungen an den Nachweis der fachlichen Unabhängigkeit von Sachverständigen und anzuhörenden Stellen, die Art und Weise der Anhörung und deren Auswertung regelt.

Der G-BA hat in seiner Sitzung am 17. Juni 2021 beschlossen, ein Stellungnahmeverfahren zur Änderung der Modulvorlage in der Anlage II zum 5. Kapitel der Verfahrensordnung einzuleiten. Gemäß dem 1. Kapitel § 8 Absatz 2 Satz 1 Buchstabe b der Verfahrensordnung des G-BA kann das Plenum im Einzelfall beschließen, dass zu Entscheidungen, bei denen kein gesetzlich eingeräumtes Stellungnahmerecht besteht, ebenfalls Stellungnahmen einzuholen sind.

Mit der geplanten Änderung der Verfahrensordnung soll eine Anpassung der Anlage II.6 (Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen) zum 5. Kapitel vorgenommen werden, welche durch Änderungen der methodischen Anforderungen an die Dossiererstellung in Verbindung mit dem bisherigen Vorgehen und den Erfahrungen des G-BA mit der Nutzenbewertung nach § 35a SGB V erforderlich geworden sind. Die Änderung betrifft in dem Abschnitt 4.3.1 (Ergebnisse randomisierter kontrollierter Studien mit dem zu bewertenden Arzneimittel) den Unterabschnitt 4.3.1.3.1 (<Endpunkt xxx> – RCT) der Anlagen II.6 zum 5. Kapitel der Verfahrensordnung. Diesbezüglich sollen Konkretisierungen zur Ergebnisdarstellung von patientenberichteten Endpunkten vorgenommen werden, wie das Vorgehen zur Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen erfolgen soll.

Zur Umsetzung der geplanten Änderung der Modulvorlage liegt ein Entwurf für die Verfahrensordnung vor, für den das Stellungnahmeverfahren nach § 35a Absatz 3 Satz 2 in Verbindung mit § 92 Absatz 3a SGB V eingeleitet wird. Nach § 92 Absatz 3a SGB V ist den Sachverständigen der medizinischen und pharmazeutischen Wissenschaft und Praxis sowie den für die Wahrnehmung der wirtschaftlichen Interessen gebildeten maßgeblichen Spitzenorganisationen der pharmazeutischen Unternehmer, den betroffenen pharmazeutischen Unternehmern, den Berufsvertretungen der Apotheker und den maßgeblichen Dachverbänden der Ärztegesellschaften der besonderen Therapierichtungen auf Bundesebene Gelegenheit zur Stellungnahme zu geben.

Die entsprechenden Entwürfe werden zu diesem Zweck dem Bundesverband der Pharmazeutischen Industrie e. V. (BPI); dem Verband Forschender Arzneimittelhersteller e. V. (vfa); dem Bundesverband der Arzneimittel-Importeure e. V. (BAI); dem Bundesverband der Arzneimittel-Hersteller e. V. (BAH); der Biotechnologie-Industrie-Organisation Deutschland e. V. (BIO Deutschland e. V.); Pro Generika e. V.; der Arzneimittelkommission der Deutschen Ärzteschaft (AkdÄ); der Arzneimittelkommission der Deutschen Zahnärzteschaft (AK-Z) c/o Bundeszahnärztekammer; der Bundesvereinigung Deutscher Apothekerverbände (ABDA); dem Deutschen Zentralverein Homöopathischer Ärzte e. V.; der Gesellschaft Anthroposophischer Ärzte e. V. und der Gesellschaft für Phytotherapie e. V. mit der Bitte um Abgabe sachverständiger Stellungnahmen der Organisationen mit Schreiben vom 22. Juni 2021 zugeleitet.

Stellungnahmen zu diesem Entwurf einschließlich Literatur sowie Literatur- bzw. Anlagenverzeichnis sind in elektronischer Form (z. B. per CD/DVD oder per E-Mail) als Word-Datei bzw. die Literatur als PDF-Dateien

bis zum 23. Juli 2021

zu richten an:

Gemeinsamer Bundesausschuss
Abteilung Arzneimittel
Gutenbergstraße 13
10587 Berlin

E-Mail: nutzenbewertung35a@g-ba.de mit Betreffzeile: „Änderung der Modulvorlage“

Betroffene pharmazeutische Unternehmen und Organisationen, die nicht Mitglieder der oben genannten Verbände sind, erhalten den Entwurf, die Technische Anlage sowie die Tragenden Gründe bei der Geschäftsstelle des G-BA.

Der Beschluss und die Tragenden Gründe können auf den Internetseiten des G-BA unter www.g-ba.de eingesehen werden.

Die mündliche Anhörung wird am 28. September 2021 um 10.00 Uhr in der Geschäftsstelle des G-BA durchgeführt. Voraussetzung für die Teilnahme an der mündlichen Anhörung ist die Abgabe einer schriftlichen Stellungnahme. Bitte melden Sie sich zeitgleich mit der Einreichung der schriftlichen Stellungnahme zu der mündlichen Anhörung an, sofern Sie an dieser teilnehmen möchten.

Berlin, den 17. Juni 2021

Gemeinsamer Bundesausschuss
gemäß § 91 SGB V

Der Vorsitzende
Prof. Hecken

Gemeinsamer
Bundesausschuss

gemäß § 91 SGB V
Unterausschuss
Arzneimittel

Besuchsadresse:
Gutenbergstr. 13
10587 Berlin

Ansprechpartner/in:
Abteilung Arzneimittel

Telefon:
030 275838210

Telefax:
030 275838205

E-Mail:
nutzenbewertung35a@g-ba.de

Internet:
www.g-ba.de

Unser Zeichen:
beh

Datum:
22. Juni 2021

Gemeinsamer Bundesausschuss, Postfach 12 06 06, 10596 Berlin

An die
Stellungnahmeberechtigten
nach §§ 35a Absatz 3 Satz 2 i.V.m. 92
Absatz 3a SGB V i.V.m. § 91 Absatz 4
Nummer 1 SGB V

Stellungnahmeverfahren zur Änderung der Verfahrensordnung nach § 35a Absatz 3 Satz 2 in Verbindung mit § 92 Absatz 3a SGB V in Verbindung mit § 91 Absatz 4 Nummer 1 SGB V

Sehr geehrte Damen und Herren,

das Plenum des Gemeinsamen Bundesausschusses (G-BA) hat in seiner Sitzung am 17. Juni 2021 beschlossen, ein Stellungnahmeverfahren zur Änderung der Verfahrensordnung nach § 35a Absatz 3a Satz 2 in Verbindung mit § 92 Absatz 3a SGB V in Verbindung mit § 91 Absatz 4 Nummer 1 SGB V einzuleiten.

Gemäß dem 1. Kapitel § 8 Absatz 2 Satz 1 Buchstabe b Verfahrensordnung des G-BA kann das Plenum im Einzelfall beschließen, dass zu Entscheidungen, bei denen kein gesetzlich eingeräumtes Stellungnahmerecht besteht, ebenfalls Stellungnahmen einzuholen sind.

Den Stellungnahmeberechtigten nach § 35a Absatz 3 Satz 2 SGB V in Verbindung mit § 92 Absatz 3a SGB V wird Gelegenheit gegeben, zu der folgenden beabsichtigten Änderung der Verfahrensordnung Stellung zu nehmen:

> 5. Kapitel der Verfahrensordnung: Änderung der Modulvorlage in der Anlage II:
>
> Anpassung der Anlage II.6 (Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen) zur Konkretisierung der Ergebnisdarstellung von patientenberichteten Endpunkten zur Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen

Stellungnehmer, die über Studiendaten verfügen, bei denen Responderanalysen im Sinne einer individuellen Minimal Important Difference (MID) vom G-BA in abgeschlossenen Nutzenbewertungen berücksichtigt wurden, werden gebeten, eine Gegenüberstellung dieser Ergebnisse mit denen einer Responseschwelle von 15% der Skalenspannweite des Instruments in das Stellungnahmeverfahren einzubringen.

Gemeinsamer Bundesausschuss

Im Rahmen Ihres Stellungnahmerechts nach § 35a Absatz 3 Satz 2 in Verbindung mit und § 92 Absatz 3a SGB V in Verbindung mit § 91 Absatz 4 Nummer 1 SGB V erhalten Sie bis zum

23. Juli 2021

Gelegenheit zur Abgabe Ihrer Stellungnahme. Später bei uns eingegangene Stellungnahmen können nicht berücksichtigt werden.

Mit Abgabe einer Stellungnahme erklären Sie sich einverstanden, dass diese in den Tragenden Gründen bzw. in der Zusammenfassenden Dokumentation wiedergegeben werden kann. Diese Dokumente werden jeweils mit Abschluss der Beratungen im Gemeinsamen Bundesausschuss erstellt und in der Regel der Öffentlichkeit via Internet zugänglich gemacht.

Ihre Stellungnahme einschließlich Literatur sowie Literatur- bzw. Anlagenverzeichnis richten Sie bitte in elektronischer Form (z. B. per CD/DVD oder per E-Mail) als Word-Datei bzw. die Literatur als PDF-Datei an:

Gemeinsamer Bundesausschuss
Abteilung Arzneimittel
Gutenbergstraße 13
10587 Berlin

E-Mail: **nutzenbewertung35a@g-ba.de** mit Betreffzeile: „Änderung der Modulvorlage“

Für Rückfragen stehen wir Ihnen gern zur Verfügung.

Mit freundlichen Grüßen

Beschluss

Gemeinsamer Bundesausschuss

des Gemeinsamen Bundesausschusses über die Einleitung eines Stellungnahmeverfahrens zur Änderung der Verfahrensordnung: Änderung der Modulvorlagen in der Anlage II zum 5. Kapitel

Vom 17. Juni 2021

Der Gemeinsame Bundesausschusses hat in seiner Sitzung am 17. Juni 2021 die Einleitung eines fakultativen Stellungnahmeverfahrens gemäß 1. Kapitel § 8 Absatz 2 Satz 1 Buchstabe b der Verfahrensordnung zur Änderung der Verfahrensordnung in der Fassung vom 18. Dezember 2008 (BAnz. Nr. 84a vom 10. Juni 2009), die zuletzt durch die Bekanntmachung des Beschlusses vom TT. Monat JJJJ (BAnz AT TT.MM.JJJJ BX) geändert worden ist, beschlossen.

Den Stellungnahmeberechtigten nach § 35a Absatz 3 Satz 2 SGB V in Verbindung mit § 92 Absatz 3a SGB V wird Gelegenheit gegeben, innerhalb einer Frist von vier Wochen zu der folgenden beabsichtigten Änderung der Verfahrensordnung Stellung zu nehmen:

I. Die Anlage II zum 5. Kapitel wird wie folgt geändert:

   Die Anlage II.6 (Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen) erhält die aus dem Anhang 1 zu diesem Beschluss ersichtliche Fassung.

II. Die Änderung der Verfahrensordnung tritt am Tag nach der Veröffentlichung im Bundesanzeiger in Kraft.

Die Tragenden Gründe zu diesem Beschluss werden auf den Internetseiten des Gemeinsamen Bundesausschusses unter www.g-ba.de veröffentlicht.

Berlin, den 17. Juni 2021

Gemeinsamer Bundesausschuss
gemäß § 91 SGB V
Der Vorsitzende

Prof. Hecken

Dokumentvorlage, Version vom 04.11.2021

Dossier zur Nutzenbewertung gemäß § 35a SGB V

<<Wirkstoff>> *(<<Handelsname>>)*

<<Pharmazeutischer Unternehmer>>

Modul 4 <<Kodierung A-Z>>

<<Anwendungsgebiet>>

Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen

Stand: <<tt.mm.jjjj>>

Inhaltsverzeichnis

Seite

Tabellenverzeichnis

Seite

Abbildungsverzeichnis

Seite

Abkürzungsverzeichnis

Abkürzung	**Bedeutung**
CONSORT	Consolidated Standards of Reporting Trials
CTCAE	Common Terminology Criteria for Adverse Events
DIMDI	Deutsches Institut für Medizinische Dokumentation
EG	Europäische Gemeinschaft
ITT	Intention to treat
MedDRA	Medical Dictionary for Regulatory Activities
MMRM	Mixed effect Model Repeat Measurement
MTC	Mixed Treatment Comparison
PT	Preferred Terms nach MedDRA
RCT	Randomized Controlled Trial
SGB	Sozialgesetzbuch
SMQs	Standardised MedDRA Queries
SOC	System Organ Class nach MedDRA
STE	Surrogate Threshold Effects
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
SUE	Schwerwiegendes UE
TREND	Transparent Reporting of Evaluations with Non-Randomized Design
UE	Unerwünschtes Ereignis
WHO	World Health Organization

4 Modul 4 – allgemeine Informationen

Modul 4 enthält folgende Angaben:

– Zusammenfassung (Abschnitt 4.1)

– Angaben zur Methodik der im Dossier präsentierten Bewertung des medizinischen Nutzens und des medizinischen Zusatznutzens (Abschnitt 4.2)

– Ergebnisse zum medizinischen Nutzen und medizinischen Zusatznutzen (Abschnitt 4.3)

– eine abschließende Bewertung der Unterlagen zum Nachweis des Zusatznutzens, einschließlich der Angabe von Patientengruppen, für die ein therapeutisch bedeutsamer Zusatznutzen besteht (Abschnitt 4.4)

– ergänzende Informationen zur Begründung der vorgelegten Unterlagen (Abschnitt 4.5)

Für jedes zu bewertende Anwendungsgebiet ist eine separate Version des vorliegenden Dokuments zu erstellen. Die Kodierung der Anwendungsgebiete ist in Modul 2 hinterlegt. Sie ist je Anwendungsgebiet einheitlich für die Module 3, 4 und 5 zu verwenden.

Im Dokument verwendete Abkürzungen sind in das Abkürzungsverzeichnis aufzunehmen. Sofern Sie für Ihre Ausführungen Tabellen und Abbildungen verwenden, sind diese im Tabellen- bzw. Abbildungsverzeichnis aufzuführen.

4.1 Zusammenfassung der Inhalte von Modul 4

Stellen Sie eine strukturierte Zusammenfassung der Inhalte von Modul 4 zur Verfügung.

Fragestellung

<< Angaben des pharmazeutischen Unternehmers >>

Datenquellen

<< Angaben des pharmazeutischen Unternehmers >>

Ein-/Ausschlusskriterien für Studien

<< Angaben des pharmazeutischen Unternehmers >>

Methoden zur Bewertung der Aussagekraft der Nachweise und zur Synthese von Ergebnissen

<< Angaben des pharmazeutischen Unternehmers >>

Ergebnisse zum medizinischen Nutzen und medizinischen Zusatznutzen

<< Angaben des pharmazeutischen Unternehmers >>

Schlussfolgerungen zum Zusatznutzen und zum therapeutisch bedeutsamen Zusatznutzen

<< Angaben des pharmazeutischen Unternehmers >>

4.2 Methodik

Abschnitt 4.2 soll die Methodik der im Dossier präsentierten Bewertung des medizinischen Nutzens und des medizinischen Zusatznutzens beschreiben. Der Abschnitt enthält Hilfestellungen für die Darstellung der Methodik sowie einige Vorgaben, die aus den internationalen Standards der evidenzbasierten Medizin abgeleitet sind. Eine Abweichung von diesen methodischen Vorgaben ist möglich, bedarf aber einer Begründung.

4.2.1 Fragestellung

Nach den internationalen Standards der evidenzbasierten Medizin soll eine Bewertung unter einer definierten Fragestellung vorgenommen werden, die mindestens folgende Komponenten enthält:

– Patientenpopulation
– Intervention
– Vergleichstherapie
– Endpunkte
– Studientypen

Unter Endpunkte sind dabei alle für die frühe Nutzenbewertung relevanten Endpunkte anzugeben (d. h. nicht nur solche, die ggf. in den relevanten Studien untersucht wurden).

Die Benennung der Vergleichstherapie in Modul 4 muss zur Auswahl der zweckmäßigen Vergleichstherapie im zugehörigen Modul 3 konsistent sein.

Geben Sie die Fragestellung der vorliegenden Aufarbeitung von Unterlagen zur Untersuchung des medizinischen Nutzens und des medizinischen Zusatznutzens des zu bewertenden Arzneimittels an. Begründen Sie Abweichungen von den oben beschriebenen Vorgaben.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.2 Kriterien für den Einschluss von Studien in die Nutzenbewertung

Die Untersuchung der in Abschnitt 4.2.1 benannten Fragestellung soll auf Basis von klinischen Studien vorgenommen werden. Für die systematische Auswahl von Studien für diese Untersuchung sollen Ein- und Ausschlusskriterien für die Studien definiert werden. Dabei ist zu beachten, dass eine Studie nicht allein deshalb ausgeschlossen werden soll, weil keine in einer Fachzeitschrift veröffentlichte Vollpublikation vorliegt. Eine Bewertung der Studie kann beispielsweise auch auf Basis eines ausführlichen Ergebnisberichts aus einem Studienregister/ einer Studienergebnisdatenbank erfolgen, während ein Kongressabstrakt allein in der Regel nicht für eine Studienbewertung ausreicht.

Benennen Sie die Ein- und Ausschlusskriterien für Studien zum medizinischen Nutzen und Zusatznutzen. Machen Sie dabei mindestens Aussagen zur Patientenpopulation, zur Intervention, zur Vergleichstherapie, zu den Endpunkten, zum Studientyp und zur Studiendauer

und begründen Sie diese. Stellen Sie die Ein- und Ausschlusskriterien zusammenfassend in einer tabellarischen Übersicht dar. Erstellen Sie dabei für unterschiedliche Themen der Recherche (z. B. unterschiedliche Fragestellungen) jeweils eine separate Übersicht.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.3 Informationsbeschaffung

In den nachfolgenden Abschnitten ist zu beschreiben, nach welcher Methodik Studien identifiziert wurden, die für die Bewertung des medizinischen Nutzens und des medizinischen Zusatznutzens in dem in diesem Dokument bewerteten Anwendungsgebiet herangezogen werden. Dies bezieht sich sowohl auf publizierte als auch auf unpublizierte Studien. Die Methodik muss dazu geeignet sein, die relevanten Studien (gemäß den in Abschnitt 4.2.2 genannten Kriterien) systematisch zu identifizieren (systematische Literaturrecherche).

4.2.3.1 Studien des pharmazeutischen Unternehmers

Für die Identifikation der Studien des pharmazeutischen Unternehmers ist keine gesonderte Beschreibung der Methodik der Informationsbeschaffung erforderlich. Die vollständige Auflistung aller Studien, die an die Zulassungsbehörde übermittelt wurden (Zulassungsstudien), sowie aller Studien, für die der pharmazeutische Unternehmer Sponsor ist oder war oder auf andere Weise finanziell beteiligt ist oder war, erfolgt in den Abschnitten 4.3.1 und 4.3.2, jeweils im Unterabschnitt „Studien des pharmazeutischen Unternehmers“. Die Darstellung soll auf Studien mit Patienten in dem Anwendungsgebiet, für das das vorliegende Dokument erstellt wird, beschränkt werden.

4.2.3.2 Bibliografische Literaturrecherche

Die Durchführung einer bibliografischen Literaturrecherche ist erforderlich, um sicherzustellen, dass ein vollständiger Studienpool in die Bewertung einfließt.

Eine bibliografische Literaturrecherche muss für RCT mit dem zu bewertenden Arzneimittel (Abschnitt 4.3.1) immer durchgeführt werden. Für indirekte Vergleiche auf Basis von RCT (Abschnitt 4.3.2.1), nicht randomisierte vergleichende Studien (Abschnitt 4.3.2.2) sowie weitere Untersuchungen (Abschnitt 4.3.2.3) muss eine bibliografische Literaturrecherche immer dann durchgeführt werden, wenn auf Basis solcher Studien der medizinische Zusatznutzen bewertet wird.

Das Datum der Recherche soll nicht mehr als 3 Monate vor dem für die Einreichung des Dossiers maßgeblichen Zeitpunkt liegen.

Die bibliografische Literaturrecherche soll mindestens in den Datenbanken MEDLINE (inklusive „in-process & other non-indexed citations) und EMBASE sowie in der Cochrane-Datenbank „Cochrane Central Register of Controlled Trials (Clinical Trials)“ durchgeführt werden. Optional kann zusätzlich eine Suche in weiteren themenspezifischen Datenbanken (z. B. CINAHL, PsycINFO etc.) durchgeführt werden.

Die Suche soll in jeder Datenbank einzeln und mit einer für die jeweilige Datenbank adaptierten Suchstrategie durchgeführt werden. Die Suchstrategien sollen jeweils in Blöcken, insbesondere getrennt nach Indikation, Intervention und ggf. Studientypen, aufgebaut werden. Wird eine Einschränkung der Strategien auf bestimmte Studientypen vorgenommen (z. B. randomisierte kontrollierte Studien), sollen aktuelle validierte Filter hierfür verwendet werden. Alle Suchstrategien sind in Anhang 4-A zu dokumentieren.

Beschreiben Sie nachfolgend für alle durchgeführten Recherchen, in welchen Datenbanken eine bibliografische Literaturrecherche durchgeführt wurde. Begründen Sie Abweichungen von den oben beschriebenen Vorgaben. Geben Sie auch an, wenn bei der Recherche generelle Einschränkungen vorgenommen wurden (z. B. Sprach- oder Jahreseinschränkungen), und begründen Sie diese.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.3.3 Suche in Studienregistern/ Studienergebnisdatenbanken

Eine Suche in öffentlich zugänglichen Studienregistern/ Studienergebnisdatenbanken ist grundsätzlich durchzuführen, um sicherzustellen, dass laufende Studien sowie abgeschlossene Studien auch von Dritten vollständig identifiziert werden und in Studienregistern / Studienergebnisdatenbanken vorliegende Informationen zu Studienmethodik und –ergebnissen in die Bewertung einfließen.

Eine Suche in Studienregistern/ Studienergebnisdatenbanken muss für RCT mit dem zu bewertenden Arzneimittel (Abschnitt 4.3.1) immer durchgeführt werden. Für indirekte Vergleiche auf Basis von RCT (Abschnitt 4.3.2.1), nicht randomisierte vergleichende Studien (Abschnitt 4.3.2.2) sowie weitere Untersuchungen (Abschnitt 4.3.2.3) muss eine Suche in Studienregistern sowie Studienergebnisdatenbanken immer dann durchgeführt werden, wenn auf Basis solcher Studien der medizinische Zusatznutzen bewertet wird.

Das Datum der Recherche soll nicht mehr als 3 Monate vor dem für die Einreichung des Dossiers maßgeblichen Zeitpunkt liegen.

Die Suche soll mindestens in den Studienregistern/ Studienergebnisdatenbanken clinicaltrials.gov (www.clinicaltrials.gov), EU Clinical Trials Register (EU-CTR, www.clinicaltrialsregister.eu), International Clinical Trials Registry Platform Search Portal (ICTRP Search Portal), Suchportal der WHO, Clinical Data Suchportal der European Medicines Agency (https://clinicaldata.ema.europa.eu) sowie dem Arzneimittel-Informationssystem (AMIS, https://www.pharmnet-bund.de/dynamic/de/arzneimittel-informationssystem/index.html) durchgeführt werden. Optional kann zusätzlich eine Suche in weiteren themenspezifischen Studienregistern / Studienergebnisdatenbanken (z. B. krankheitsspezifische Studienregister oder Studienregister einzelner pharmazeutischer Unternehmen) durchgeführt werden. Die Suche in Studienregistern/ Studienergebnisdatenbanken anderer pharmazeutischer Unternehmer ist insbesondere bei

indirekten Vergleichen sinnvoll, wenn Studien zu anderen Arzneimitteln identifiziert werden müssen.

Die Suche soll in jedem Studienregister/ Studienergebnisdatenbank einzeln und mit einer für das jeweilige Studienregister/ Studienergebnisdatenbank adaptierten Suchstrategie durchgeführt werden. Die Suche soll abgeschlossene, abgebrochene und laufende Studien erfassen. Alle Suchstrategien sind in Anhang 4-B zu dokumentieren.

Für Clinical Data (Suchportal der European Medicines Agency) und das Arzneimittel-Informationssystem (AMIS) genügt hingegen die Suche nach Einträgen mit Ergebnisberichten zu Studien, die bereits anderweitig (z.B. über die bibliografische Literaturrecherche und Studienregistersuche) identifiziert wurden. Eine Dokumentation der zugehörigen Suchstrategie ist nicht erforderlich.

Beschreiben Sie nachfolgend für alle durchgeführten Recherchen, in welchen Studienregistern/ Studienergebnisdatenbanken die Suche durchgeführt wurde. Begründen Sie dabei Abweichungen von den oben beschriebenen Vorgaben. Geben Sie auch an, wenn bei der Recherche generelle Einschränkungen vorgenommen wurden (z. B. Jahreseinschränkungen), und begründen Sie diese.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.3.4 Suche auf der Internetseite des G-BA

Die Internetseite des G-BA ist grundsätzlich zu durchsuchen, um sicherzustellen, dass alle vorliegenden Daten zu Studienmethodik und –ergebnissen von relevanten Studien in die Bewertung einfließen.

Auf der Internetseite des G-BA werden Dokumente zur frühen Nutzenbewertung nach §35a SGB V veröffentlicht. Diese enthalten teilweise anderweitig nicht veröffentlichte Daten zu Studienmethodik und –ergebnissen[1]. Solche Daten sind dabei insbesondere in den Modulen 4 der Dossiers pharmazeutischer Unternehmer, in IQWiG-Nutzenbewertungen sowie dem Beschluss des G-BA einschließlich der Tragenden Gründe und der Zusammenfassenden Dokumentation zu erwarten.

Die Suche auf der Internetseite des G-BA muss für RCT mit dem zu bewertenden Arzneimittel (Abschnitt 4.3.1) immer durchgeführt werden. Für indirekte Vergleiche auf Basis von RCT (Abschnitt 4.3.2.1), nicht randomisierte vergleichende Studien (Abschnitt 4.3.2.2) sowie weitere Untersuchungen (Abschnitt 4.3.2.3) muss eine Suche auf der G-BA Internetseite immer dann durchgeführt werden, wenn auf Basis solcher Studien der medizinische Zusatznutzen

[1] Köhler M, Haag S, Biester K, Brockhaus AC, McGauran N, Grouven U, Kölsch H, Seay U, Hörn H, Moritz G, Staeck K, Wieseler B. Information on new drugs at market entry: retrospective analysis of health technology assessment reports, journal publications, and registry reports. BMJ 2015;350:h796

bewertet wird. Die Suche ist dann sowohl für das zu bewertende Arzneimittel als auch für die zweckmäßige Vergleichstherapie durchzuführen. Es genügt die Suche nach Einträgen zu Studien, die bereits anderweitig (z.B. über die bibliografische Literaturrecherche und Studienregistersuche) identifiziert wurden. Eine Dokumentation der zugehörigen Suchstrategie ist nicht erforderlich.

Das Datum der Recherche soll nicht mehr als 3 Monate vor dem für die Einreichung des Dossiers maßgeblichen Zeitpunkt liegen.

Beschreiben Sie nachfolgend das Vorgehen für die Suche. Benennen Sie die Wirkstoffe und die auf der Internetseite des G-BA genannten zugehörigen Vorgangsnummern, zu denen Sie eine Suche durchgeführt haben.

Begründen Sie Abweichungen von den oben beschriebenen Vorgaben.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.3.5 Selektion relevanter Studien

Beschreiben Sie das Vorgehen bei der Selektion relevanter Studien aus dem Ergebnis der in den Abschnitten 4.2.3.2, 4.2.3.3 und 4.2.3.4 beschriebenen Rechercheschritte. Begründen Sie das Vorgehen, falls die Selektion nicht von zwei Personen unabhängig voneinander durchgeführt wurde.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.4 Bewertung der Aussagekraft der Nachweise

Zur Bewertung der Aussagekraft der im Dossier vorgelegten Nachweise sollen Verzerrungsaspekte der Ergebnisse für jede eingeschlossene Studie beschrieben werden, und zwar separat für jeden patientenrelevanten Endpunkt. Dazu sollen insbesondere folgende endpunktübergreifende (A) und endpunktspezifische (B) Aspekte systematisch extrahiert werden (zur weiteren Erläuterung der einzelnen Aspekte siehe Bewertungsbogen in Anhang 4-F):

A: Verzerrungsaspekte der Ergebnisse auf Studienebene

– Erzeugung der Randomisierungssequenz *(bei randomisierten Studien)*
– Verdeckung der Gruppenzuteilung *(bei randomisierten Studien)*
– zeitliche Parallelität der Gruppen *(bei nicht randomisierten vergleichenden Studien)*
– Vergleichbarkeit der Gruppen bzw. Berücksichtigung prognostisch relevanter Faktoren *(bei nicht randomisierten vergleichenden Studien)*
– Verblindung des Patienten sowie der behandelnden Personen

– ergebnisgesteuerte Berichterstattung
– sonstige Aspekte

B: Verzerrungsaspekte der Ergebnisse auf Endpunktebene

– Verblindung der Endpunkterheber
– Umsetzung des ITT-Prinzips
– ergebnisgesteuerte Berichterstattung
– sonstige Aspekte

Für randomisierte Studien soll darüber hinaus das Verzerrungspotenzial bewertet und als „niedrig“ oder „hoch“ eingestuft werden. Ein niedriges Verzerrungspotenzial liegt dann vor, wenn mit großer Wahrscheinlichkeit ausgeschlossen werden kann, dass die Ergebnisse relevant verzerrt sind. Unter einer relevanten Verzerrung ist zu verstehen, dass sich die Ergebnisse bei Behebung der verzerrenden Aspekte in ihrer Grundaussage verändern würden.

Eine zusammenfassende Bewertung der Verzerrungsaspekte soll nicht für nicht randomisierte Studien erfolgen.

Für die Bewertung eines Endpunkts soll für randomisierte Studien zunächst das Verzerrungspotenzial endpunktübergreifend anhand der unter A aufgeführten Aspekte als „niedrig“ oder „hoch“ eingestuft werden. Falls diese Einstufung als „hoch“ erfolgt, soll das Verzerrungspotenzial für den Endpunkt in der Regel auch als „hoch“ bewertet werden, Abweichungen hiervon sind zu begründen. Ansonsten sollen die unter B genannten endpunktspezifischen Aspekte Berücksichtigung finden.

Eine Einstufung des Verzerrungspotenzials des Ergebnisses für einen Endpunkt als „hoch“ soll nicht zum Ausschluss der Daten führen. Die Klassifizierung soll vielmehr der Diskussion heterogener Studienergebnisse und der Einschätzung der Aussagekraft der Nachweise dienen. Für nicht randomisierte Studien können für solche Diskussionen einzelne Verzerrungsaspekte herangezogen werden.

Beschreiben Sie die für die Bewertung der Verzerrungsaspekte und des Verzerrungspotenzials eingesetzte Methodik. Begründen Sie, wenn Sie von der oben beschriebenen Methodik abweichen.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.5 Informationssynthese und -analyse

4.2.5.1 Beschreibung des Designs und der Methodik der eingeschlossenen Studien

Das Design und die Methodik der eingeschlossenen Studien soll in den Abschnitten 4.3.1 und 4.3.2, jeweils in den Unterabschnitten „Charakteristika der in die Bewertung eingeschlossenen Studien“ und den dazugehörigen Anhängen, dargestellt werden. Die Darstellung der Studien

soll für randomisierte kontrollierte Studien mindestens die Anforderungen des CONSORT-Statements erfüllen (Items 2b bis 14, Informationen aus dem CONSORT-Flow-Chart)[2]. Die Darstellung nicht randomisierter Interventionsstudien und epidemiologischer Beobachtungsstudien soll mindestens den Anforderungen des TREND-[3] bzw. STROBE-Statements[4] folgen. Design und Methodik weiterer Untersuchungen sollen gemäß den verfügbaren Standards dargestellt werden.

Beschreiben Sie, nach welchen Standards und mit welchen Informationen (Items) Sie das Design und die Methodik der eingeschlossenen Studien in Modul 4 dargestellt haben. Begründen Sie Abweichungen von den oben beschriebenen Vorgaben.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.5.2 Gegenüberstellung der Ergebnisse der Einzelstudien

Die Ergebnisse der einzelnen Studien sollen in den Abschnitten 4.3.1 und 4.3.2 in den entsprechenden Unterabschnitten zunächst für jede eingeschlossene Studie separat dargestellt werden. Die Darstellung soll die Charakteristika der Studienpopulationen sowie die Ergebnisse zu allen in den eingeschlossenen Studien berichteten patientenrelevanten Endpunkten (Verbesserung des Gesundheitszustands, Verkürzung der Krankheitsdauer, Verlängerung des Überlebens, Verringerung von Nebenwirkungen, Verbesserung der Lebensqualität) umfassen. Anforderungen an die Darstellung werden in den Unterabschnitten beschrieben.

Benennen Sie die Patientencharakteristika und patientenrelevanten Endpunkte, die in den relevanten Studien erhoben wurden. Begründen Sie, wenn Sie von den oben benannten Vorgaben abgewichen sind. Beschreiben Sie für jeden Endpunkt, warum Sie ihn als patientenrelevant einstufen, und machen Sie Angaben zur Validität des Endpunkts (z. B. zur Validierung der eingesetzten Fragebögen). Geben Sie für den jeweiligen Endpunkt an, ob unterschiedliche Operationalisierungen innerhalb der Studien und zwischen den Studien verwendet wurden. Benennen Sie die für die Bewertung herangezogene(n) Operationalisierung(en) und begründen Sie die Auswahl. Beachten Sie bei der Berücksichtigung von Surrogatendpunkten Abschnitt 4.5.4.

Sofern zur Berechnung von Ergebnissen von Standardverfahren und –software abgewichen wird (insbesondere beim Einsatz spezieller Software oder individueller Programmierung), sind

[2] Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c332.
[3] Des Jarlais DC, Lyles C, Crepaz N. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Publ Health 2004; 94(3): 361-366.
[4] Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007; 147(8): 573-577.

die Berechnungsschritte und ggf. verwendete Software explizit abzubilden. Insbesondere der Programmcode ist in lesbarer Form anzugeben.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.5.3 Meta-Analysen

Sofern mehrere Studien vorliegen, sollen diese in einer Meta-Analyse quantitativ zusammengefasst werden, wenn die Studien aus medizinischen (z. B. Patientengruppen) und methodischen (z.B. Studiendesign) Gründen ausreichend vergleichbar sind. Es ist jeweils zu begründen, warum eine Meta-Analyse durchgeführt wurde oder warum eine Meta-Analyse nicht durchgeführt wurde bzw. warum einzelne Studien ggf. nicht in die Meta-Analyse einbezogen wurden. Für Meta-Analysen soll die im Folgenden beschriebene Methodik eingesetzt werden.

Für die statistische Auswertung sollen primär die Ergebnisse aus Intention-to-treat-Analysen, so wie sie in den vorliegenden Dokumenten beschrieben sind, verwendet werden. Die Meta-Analysen sollen in der Regel auf Basis von Modellen mit zufälligen Effekten nach der Knapp-Hartung-Methode mit der Paule-Mandel-Methode zur Heterogenitätsschätzung[5] erfolgen. Im Fall von sehr wenigen Studien ist die Heterogenität nicht verlässlich schätzbar. Liegen daher weniger als 5 Studien vor, ist auch die Anwendung eines Modells mit festem Effekt oder eine qualitative Zusammenfassung in Betracht zu ziehen. Kontextabhängig können auch alternative Verfahren wie z. B. Bayes'sche Verfahren oder Methoden aus dem Bereich der generalisierten linearen Modelle in Erwägung gezogen werden. Falls die für eine Meta-Analyse notwendigen Schätzer für Lage und Streuung in den Studienunterlagen nicht vorliegen, sollen diese nach Möglichkeit aus den vorhandenen Informationen eigenständig berechnet beziehungsweise näherungsweise bestimmt werden.

Für kontinuierliche Variablen soll die Mittelwertdifferenz, gegebenenfalls standardisiert mittels Hedges' g, als Effektmaß eingesetzt werden. Bei binären Variablen sollen Meta-Analysen primär sowohl anhand des Odds Ratios als auch des Relativen Risikos durchgeführt werden. In begründeten Ausnahmefällen können auch andere Effektmaße zum Einsatz kommen. Bei kategorialen Variablen soll ein geeignetes Effektmaß in Abhängigkeit vom konkreten Endpunkt und den verfügbaren Daten verwendet[6] werden.

Die Effektschätzer und Konfidenzintervalle aus den Studien sollen mittels Forest Plots zusammenfassend dargestellt werden. Anschließend soll die Einschätzung einer möglichen Heterogenität der Studienergebnisse anhand geeigneter statistische Maße auf Vorliegen von

[5] Veroniki AA, Jackson D, Viechtbauer W, Bender R, Knapp G, Kuss O et al. Recommendations for quantifying the uncertainty in the summary intervention effect and estimating the between-study heterogeneity variance in random-effects meta-analysis. Cochrane Database Syst Rev 2015: 25-27.
[6] Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S (Ed). Cochrane handbook for systematic reviews of interventions. Chichester: Wiley; 2008. S. 243-296.

Heterogenität[7, 5] erfolgen. Die Heterogenitätsmaße sind unabhängig von dem Ergebnis der Untersuchung auf Heterogenität immer anzugeben. Ist die Heterogenität der Studienergebnisse nicht bedeutsam (z. B. p-Wert für Heterogenitätsstatistik $\geq 0{,}05$), soll der gemeinsame (gepoolte) Effekt inklusive Konfidenzintervall dargestellt werden. Bei bedeutsamer Heterogenität sollen die Ergebnisse nur in begründeten Ausnahmefällen gepoolt werden. Außerdem soll untersucht werden, welche Faktoren diese Heterogenität möglicherweise erklären könnten. Dazu zählen methodische Faktoren (siehe Abschnitt 4.2.5.4) und klinische Faktoren, sogenannte Effektmodifikatoren (siehe Abschnitt 4.2.5.5).

Beschreiben Sie die für Meta-Analysen eingesetzte Methodik. Begründen Sie, wenn Sie von der oben beschriebenen Methodik abweichen.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.5.4 Sensitivitätsanalysen

Zur Einschätzung der Robustheit der Ergebnisse sollen Sensitivitätsanalysen hinsichtlich methodischer Faktoren durchgeführt werden. Die methodischen Faktoren bilden sich aus den im Rahmen der Informationsbeschaffung und -bewertung getroffenen Entscheidungen, zum Beispiel die Festlegung von Cut-off-Werten für Erhebungszeitpunkte oder die Wahl des Effektmaßes. Insbesondere die Einstufung des Verzerrungspotenzials der Ergebnisse in die Kategorien „hoch“ und „niedrig“ soll für Sensitivitätsanalysen verwendet werden.

Das Ergebnis der Sensitivitätsanalysen kann die Einschätzung der Aussagekraft der Nachweise beeinflussen.

Begründen Sie die durchgeführten Sensitivitätsanalysen oder den Verzicht auf Sensitivitätsanalysen. Beschreiben Sie die für Sensitivitätsanalysen eingesetzte Methodik. Begründen Sie, wenn Sie von der oben beschriebenen Methodik abweichen.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.5.5 Subgruppenmerkmale und andere Effektmodifikatoren

Die Ergebnisse sollen hinsichtlich potenzieller Effektmodifikatoren, das heißt klinischer Faktoren, die die Effekte beeinflussen, untersucht werden. Dies können beispielsweise direkte Patientencharakteristika (Subgruppenmerkmale) sowie Spezifika der Behandlungen (z. B. die Dosis) sein. Im Gegensatz zu den in Abschnitt 4.2.5.4 beschriebenen methodischen Faktoren für Sensitivitätsanalysen besteht hier das Ziel, mögliche Effektunterschiede zwischen Patientengruppen und Behandlungsspezifika aufzudecken. Eine potenzielle Effektmodifikation

[7] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327(7414):557-560.

soll anhand von Homogenitäts- bzw. Interaktionstests oder von Interaktionstermen aus Regressionsanalysen (mit Angabe von entsprechenden Standardfehlern) untersucht werden. Subgruppenanalysen auf der Basis individueller Patientendaten haben in der Regel eine größere Ergebnissicherheit als solche auf Basis von Meta-Regressionen oder Meta-Analysen unter Kategorisierung der Studien bezüglich der möglichen Effektmodifikatoren, sie sind deshalb zu bevorzugen. Es sollen, soweit sinnvoll, folgende Faktoren bezüglich einer möglichen Effektmodifikation berücksichtigt werden:

– Geschlecht
– Alter
– Krankheitsschwere bzw. –stadium
– Zentrums- und Ländereffekte

Sollten sich aus den verfügbaren Informationen Anzeichen für weitere mögliche Effektmodifikatoren ergeben, können diese ebenfalls begründet einbezogen werden. Die Ergebnisse von in Studien a priori geplanten und im Studienprotokoll festgelegten Subgruppenanalysen für patientenrelevante Endpunkte sind immer darzustellen (zu ergänzenden Kriterien zur Darstellung siehe Abschnitt 4.3.1.3.2).

Bei Identifizierung möglicher Effektmodifikatoren kann gegebenenfalls eine Präzisierung der aus den für die Gesamtgruppe beobachteten Effekten abgeleiteten Aussagen erfolgen. Ergebnisse von Subgruppenanalysen können die Identifizierung von Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen unterstützen.

Benennen Sie die durchgeführten Subgruppenanalysen. Begründen Sie die Wahl von Trennpunkten, wenn quantitative Merkmale kategorisiert werden. Verwenden Sie dabei nach Möglichkeit die in dem jeweiligen Gebiet gebräuchlichen Einteilungen und begründen Sie etwaige Abweichungen. Begründen Sie die durchgeführten Subgruppenanalysen bzw. die Untersuchung von Effektmodifikatoren oder den Verzicht auf solche Analysen. Beschreiben Sie die für diese Analysen eingesetzte Methodik. Begründen Sie, wenn Sie von der oben beschriebenen Methodik abweichen.

<< Angaben des pharmazeutischen Unternehmers >>

4.2.5.6 Indirekte Vergleiche

Zurzeit sind international Methoden in der Entwicklung, um indirekte Vergleiche zu ermöglichen. Es besteht dabei internationaler Konsens, dass Vergleiche einzelner Behandlungsgruppen aus verschiedenen Studien ohne Bezug zu einem gemeinsamen Komparator (häufig als nicht adjustierte indirekte Vergleiche bezeichnet) regelhaft keine valide

Analysemethode darstellen[8]. Eine Ausnahme kann das Vorliegen von dramatischen Effekten sein. An Stelle von nicht adjustierten indirekten Vergleichen sollen je nach Datenlage einfache adjustierte indirekte Vergleiche[9] oder komplexere Netzwerk-Meta-Analysen (auch als „Mixed Treatment Comparison [MTC] Meta-Analysen“ oder „Multiple Treatment Meta-Analysen“ bezeichnet) für den simultanen Vergleich von mehr als zwei Therapien unter Berücksichtigung sowohl direkter als auch indirekter Vergleiche berechnet werden. Aktuelle Verfahren wurden beispielsweise von Lu und Ades (2004)[10] und Rücker (2012)[11] vorgestellt.

Alle Verfahren für indirekte Vergleiche gehen im Prinzip von den gleichen zentralen Annahmen aus. Hierbei handelt es sich um die Annahmen der Ähnlichkeit der eingeschlossenen Studien, der Homogenität der paarweisen Vergleiche und der Konsistenz zwischen direkter und indirekter Evidenz innerhalb des zu analysierenden Netzwerkes. Als Inkonsistenz wird dabei die Diskrepanz zwischen dem Ergebnis eines direkten und eines oder mehreren indirekten Vergleichen verstanden, die nicht mehr nur durch Zufallsfehler oder Heterogenität erklärbar ist[12].

Das Ergebnis eines indirekten Vergleichs kann maßgeblich von der Auswahl des Brückenkomparators bzw. der Brückenkomparatoren abhängen. Als Brückenkomparatoren sind dabei insbesondere Interventionen zu berücksichtigen, für die sowohl zum bewertenden Arzneimittel als auch zur zweckmäßigen Vergleichstherapie mindestens eine direkt vergleichende Studie vorliegt (Brückenkomparatoren ersten Grades).

Insgesamt ist es notwendig, die zugrunde liegende Methodik für alle relevanten Endpunkte genau und reproduzierbar zu beschreiben und die zentralen Annahmen zu untersuchen[13, 14, 15]

[8] Bender R, Schwenke C, Schmoor C, Hauschke D. Stellenwert von Ergebnissen aus indirekten Vergleichen - Gemeinsame Stellungnahme von IQWiG, GMDS und IBS-DR [online]. [Zugriff: 31.10.2016]. URL: http://www.gmds.de/pdf/publikationen/stellungnahmen/120202_IQWIG_GMDS_IBS_DR.pdf.

[9] Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997; 50(6): 683-691.

[10] Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004; 23(20): 3105-3124.

[11] Rücker G. Network meta-analysis, electrical networks and graph theory. Res Synth Methods 2012; 3(4): 312-324.

[12] Schöttker B, Lühmann D, Boulkhemair D, Raspe H. Indirekte Vergleiche von Therapieverfahren. Schriftenreihe Health Technology Assessment Band 88, DIMDI, Köln, 2009.

[13] Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DJ. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ 2009; 338: b1147.

[14] Song F, Xiong T, Parekh-Bhurke S, Loke YK, Sutton AJ, Eastwood AJ et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study BMJ 2011; 343 :d4909

[15] Donegan S, Williamson P, D'Alessandro U, Tudur Smith C. Assessing key assumptions of network meta-analysis: a review of methods. Res Synth Methods 2013; 4(4): 291-323.

Beschreiben Sie detailliert und vollständig die zugrunde liegende Methodik des indirekten Vergleichs. Dabei sind mindestens folgende Angaben notwendig:

- *Benennung aller potentiellen Brückenkomparatoren ersten Grades und ggf. Begründung für die Auswahl.*

- *Genaue Spezifikation des statistischen Modells inklusive aller Modellannahmen. Bei Verwendung eines Bayes'schen Modells sind dabei auch die angenommenen A-priori-Verteilungen (falls informative Verteilungen verwendet werden, mit Begründung), die Anzahl der Markov-Ketten, die Art der Untersuchung der Konvergenz der Markov-Ketten und deren Startwerte und Länge zu spezifizieren.*

- *Art der Prüfung der Ähnlichkeit der eingeschlossenen Studien.*

- *Art der Prüfung der Homogenität der Ergebnisse direkter paarweiser Vergleiche.*

- *Art der Prüfung der Konsistenzannahme im Netzwerk.*

- *Bilden Sie den Code des Computerprogramms inklusive der einzulesenden Daten in lesbarer Form ab und geben Sie an, welche Software Sie zur Berechnung eingesetzt haben (ggf. inklusive Spezifizierung von Modulen, Prozeduren, Packages etc.; siehe auch Modul 5 zur Ablage des Programmcodes).*

- *Art und Umfang von Sensitivitätsanalysen.*

<< Angaben des pharmazeutischen Unternehmers >>

4.3 Ergebnisse zum medizinischen Nutzen und zum medizinischen Zusatznutzen

In den nachfolgenden Abschnitten sind die Ergebnisse zum medizinischen Nutzen und zum medizinischen Zusatznutzen zu beschreiben. Abschnitt 4.3.1 enthält dabei die Ergebnisse aus randomisierten kontrollierten Studien, die mit dem zu bewertenden Arzneimittel durchgeführt wurden (Evidenzstufen Ia/Ib).

Abschnitt 4.3.2 enthält weitere Unterlagen anderer Evidenzstufen, sofern diese aus Sicht des pharmazeutischen Unternehmers zum Nachweis des Zusatznutzens erforderlich sind. Diese Unterlagen teilen sich wie folgt auf:

– Randomisierte, kontrollierte Studien für einen indirekten Vergleich mit der zweckmäßigen Vergleichstherapie, sofern keine direkten Vergleichsstudien mit der zweckmäßigen Vergleichstherapie vorliegen oder diese keine ausreichenden Aussagen über den Zusatznutzen zulassen (Abschnitt 4.3.2.1)
– Nicht randomisierte vergleichende Studien (Abschnitt 4.3.2.2)
– Weitere Untersuchungen (Abschnitt 4.3.2.3)

Falls für die Bewertung des Zusatznutzens mehrere Komparatoren (z.B. Wirkstoffe) herangezogen werden, sind die Aussagen zum Zusatznutzen primär gegenüber der Gesamtheit der gewählten Komparatoren durchzuführen (z. B. basierend auf Meta-Analysen unter gemeinsamer Betrachtung aller direkt vergleichender Studien). Spezifische methodische Argumente, die gegen eine gemeinsame Analyse sprechen (z. B. statistische oder inhaltliche Heterogenität), sind davon unbenommen. Eine zusammenfassende Aussage zum Zusatznutzen gegenüber der zweckmäßigen Vergleichstherapie ist in jedem Fall erforderlich.

4.3.1 Ergebnisse randomisierter kontrollierter Studien mit dem zu bewertenden Arzneimittel

4.3.1.1 Ergebnis der Informationsbeschaffung – RCT mit dem zu bewertenden Arzneimittel

4.3.1.1.1 Studien des pharmazeutischen Unternehmers

Nachfolgend sollen alle Studien (RCT), die an die Zulassungsbehörde übermittelt wurden (Zulassungsstudien), sowie alle Studien (RCT), für die der pharmazeutische Unternehmer Sponsor ist oder war oder auf andere Weise finanziell beteiligt ist oder war, benannt werden. Beachten Sie dabei folgende Konkretisierungen:

- *Es sollen alle RCT, die der Zulassungsbehörde im Zulassungsdossier übermittelt wurden und deren Studienberichte im Abschnitt 5.3.5 des Zulassungsdossiers enthalten sind, aufgeführt werden. Darüber hinaus sollen alle RCT, für die der pharmazeutische Unternehmer Sponsor ist oder war oder auf andere Weise finanziell beteiligt ist oder war, aufgeführt werden.*

- *Benennen Sie in der nachfolgenden Tabelle nur solche RCT, die ganz oder teilweise innerhalb des in diesem Dokument beschriebenen Anwendungsgebiets durchgeführt wurden. Fügen Sie dabei für jede Studie eine neue Zeile ein.*

Folgende Informationen sind in der Tabelle darzulegen: Studienbezeichnung, Angabe „Zulassungsstudie ja/nein", Angabe über die Beteiligung (Sponsor ja/nein), Studienstatus (abgeschlossen, abgebrochen, laufend), Studiendauer, Angabe zu geplanten und durchgeführten Datenschnitten und Therapiearme. Orientieren Sie sich dabei an der beispielhaften Angabe in der ersten Tabellenzeile.

Tabelle 4-1: Liste der Studien des pharmazeutischen Unternehmers – RCT mit dem zu bewertenden Arzneimittel

Studie	**Zulassungsstudie (ja/nein)**	**Sponsor (ja/nein)**	**Status (abgeschlossen / abgebrochen / laufend)**	**Studiendauer ggf. Datenschnitt**	**Therapiearme**
<Studie 1>	ja	ja	abgeschlossen	12 Monate	Medikament A, Medikament B, Placebo

Geben Sie an, welchen Stand die Information in Tabelle 4-1 hat, d. h. zu welchem Datum der Studienstatus abgebildet wird. Das Datum des Studienstatus soll nicht mehr als 3 Monate vor dem für die Einreichung des Dossiers maßgeblichen Zeitpunkt liegen.

<< Angaben des pharmazeutischen Unternehmers >>

Geben Sie in der nachfolgenden Tabelle an, welche der in Tabelle 4-1 genannten Studien nicht für die Nutzenbewertung herangezogen wurden. Begründen Sie dabei jeweils die Nichtberücksichtigung. Fügen Sie für jede Studie eine neue Zeile ein.

Tabelle 4-2: Studien des pharmazeutischen Unternehmers, die nicht für die Nutzenbewertung herangezogen wurden – RCT mit dem zu bewertenden Arzneimittel

Studienbezeichnung	**Begründung für die Nichtberücksichtigung der Studie**

4.3.1.1.2 Studien aus der bibliografischen Literaturrecherche

Beschreiben Sie nachfolgend das Ergebnis der bibliografischen Literaturrecherche. Illustrieren Sie den Selektionsprozess und das Ergebnis der Selektion mit einem Flussdiagramm. Geben Sie dabei an, wie viele Treffer sich insgesamt (d. h. über alle durchsuchten Datenbanken) aus der bibliografischen Literaturrecherche ergeben haben, wie viele Treffer sich nach Entfernung von Dubletten ergeben haben, wie viele Treffer nach Sichtung von Titel und, sofern vorhanden, Abstract als nicht relevant angesehen wurden, wie viele Treffer im Volltext gesichtet wurden, wie viele der im Volltext gesichteten Treffer nicht relevant waren (mit Angabe der Ausschlussgründe) und wie viele relevante Treffer verblieben. Geben Sie zu den relevanten Treffern an, wie vielen Einzelstudien diese zuzuordnen sind. Listen Sie die im Volltext gesichteten und ausgeschlossenen Dokumente unter Nennung des Ausschlussgrunds in Anhang 4-C.

[Anmerkung: „Relevanz“ bezieht sich in diesem Zusammenhang auf die im Abschnitt 4.2.2 genannten Kriterien für den Einschluss von Studien in die Nutzenbewertung.]

Geben Sie im Flussdiagramm auch das Datum der Recherche an. Die Recherche soll nicht mehr als 3 Monate vor dem für die Einreichung des Dossiers maßgeblichen Zeitpunkt liegen.

Orientieren Sie sich bei der Erstellung des Flussdiagramms an dem nachfolgenden Beispiel.

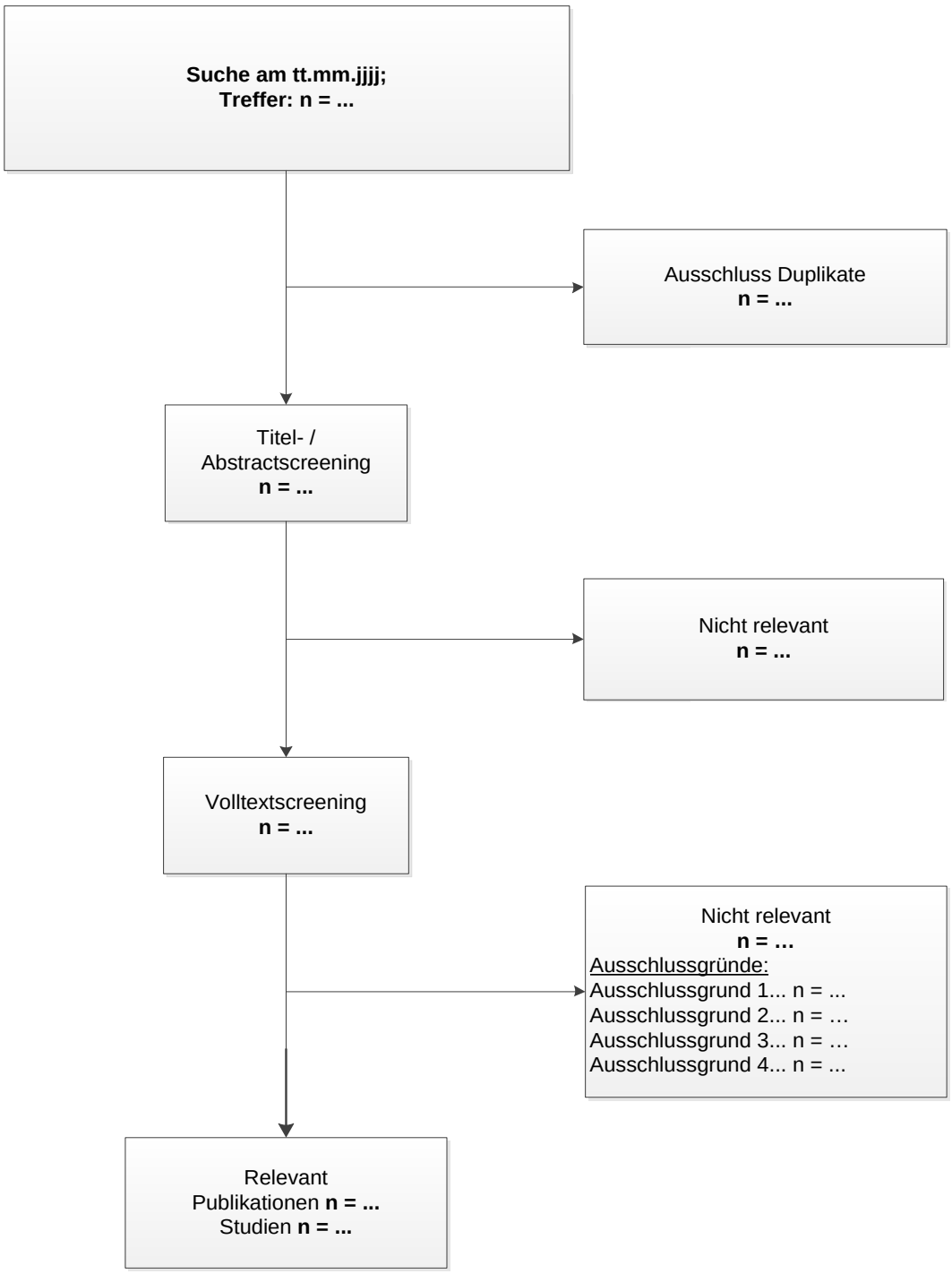

Abbildung 1: Flussdiagramm der bibliografischen Literaturrecherche – Suche nach randomisierten kontrollierten Studien mit dem zu bewertenden Arzneimittel

4.3.1.1.3 Studien aus der Suche in Studienregistern/ Studienergebnisdatenbanken

Beschreiben Sie in der nachfolgenden Tabelle alle relevanten Studien, die durch die Suche in Studienregistern/ Studienergebnisdatenbanken identifiziert wurden. Geben Sie dabei an, in welchem Studienregister / Studienergebnisdatenbank die Studie identifiziert wurde und welche

Dokumente dort zur Studie jeweils hinterlegt sind (z. B. Studienregistereintrag, Bericht über Studienergebnisse etc.). Geben Sie auch an, ob die Studie in der Liste der Studien des pharmazeutischen Unternehmers enthalten ist (siehe Tabelle 4-1) und ob die Studie auch durch die bibliografische Literaturrecherche identifiziert wurde. Fügen Sie für jede Studie eine neue Zeile ein. Listen Sie die ausgeschlossenen Studien unter Nennung des Ausschlussgrunds in Anhang 4-D.

[Anmerkung: „Relevanz" bezieht sich in diesem Zusammenhang auf die im Abschnitt 4.2.2 genannten Kriterien für den Einschluss von Studien in die Nutzenbewertung.]

Orientieren Sie sich bei Ihren Angaben an der beispielhaften ersten Tabellenzeile.

Tabelle 4-3: Relevante Studien (auch laufende Studien) aus der Suche in Studienregistern / Studienergebnisdatenbanken – RCT mit dem zu bewertenden Arzneimittel

Studie	**Identifikationsorte (Name des Studienregisters/ der Studienergebnisdatenbank und Angabe der Zitate[a])**	**Studie in Liste der Studien des pharmazeutischen Unternehmers enthalten (ja/nein)**	**Studie durch bibliografische Literaturrecherche identifiziert (ja/nein)**	**Status (abgeschlossen/ abgebrochen/ laufend)**
<Studie 1>	NCT 12345 [6, 7]	ja	nein	abgeschlossen
	EudraCT 1223456 [8, 9]			
a: Zitat des Studienregistereintrags, die Studienregisternummer (NCT-Nummer, EudraCT-Nummer) sowie, falls vorhanden, der im Studienregister/in der Studienergebnisdatenbank aufgelisteten Berichte über Studiendesign und/oder -ergebnisse.				

Geben Sie an, welchen Stand die Information in Tabelle 4-3 hat, d. h. zu welchem Datum die Recherche durchgeführt wurde. Das Datum der Recherche soll nicht mehr als 3 Monate vor dem für die Einreichung des Dossiers maßgeblichen Zeitpunkt liegen.

<< Angaben des pharmazeutischen Unternehmers >>

4.3.1.1.4 Studien aus der Suche auf der Internetseite des G-BA

Beschreiben Sie in der nachfolgenden Tabelle alle relevanten Studien, die durch die Sichtung der Internetseite des G-BA identifiziert wurden. Geben Sie dabei an, welche Dokumente dort hinterlegt sind (z. B. Dossier eines anderen pharmazeutischen Unternehmers, IQWiG Nutzenbewertung). Geben Sie auch an, ob die Studie in der Liste der Studien des pharmazeutischen Unternehmers enthalten ist (siehe Tabelle 4-1) und ob die Studie auch durch die bibliografische Literaturrecherche bzw. Suche in Studienregistern/ Studienergebnisdatenbank identifiziert wurde. Fügen Sie für jede Studie eine neue Zeile ein.

[Anmerkung: „Relevanz" bezieht sich in diesem Zusammenhang auf die im Abschnitt 4.2.2 genannten Kriterien für den Einschluss von Studien in die Nutzenbewertung.]

Orientieren Sie sich bei Ihren Angaben an der beispielhaften ersten Tabellenzeile.

Tabelle 4-4: Relevante Studien aus der Suche auf der Internetseite des G-BA – RCT mit dem zu bewertenden Arzneimittel

Studie	**Relevante Quellen**[a]	**Studie in Liste der Studien des pharmazeutischen Unternehmers enthalten (ja/nein)**	**Studie durch bibliografische Literaturrecherche identifiziert (ja/nein)**	**Studie durch Suche in Studienregistern / Studienergebnis datenbanken identifiziert (ja/nein)**
<Studie 1>	Dossier, Modul 4 (Vorgangsnummer 2013-10-01-D-076) [8] IQWiG Nutzenbewertung (A13-35) [9]	Ja	Nein	Ja
a: Quellen aus der Suche auf der Internetseite des G-BA				

Geben Sie an, welchen Stand die Information in Tabelle 4-4 hat, d. h. zu welchem Datum die Recherche durchgeführt wurde. Das Datum der Recherche soll nicht mehr als 3 Monate vor dem für die Einreichung des Dossiers maßgeblichen Zeitpunkt liegen.

<< Angaben des pharmazeutischen Unternehmers >>

4.3.1.1.5 Resultierender Studienpool: RCT mit dem zu bewertenden Arzneimittel

Benennen Sie in der nachfolgenden Tabelle den aus den verschiedenen Suchschritten (Abschnitte 4.3.1.1.1, 4.3.1.1.2, 4.3.1.1.3 und 4.3.1.1.4) resultierenden Pool relevanter Studien (exklusive laufender Studien) für das zu bewertende Arzneimittel, auch im direkten Vergleich zur zweckmäßigen Vergleichstherapie. Führen Sie außerdem alle relevanten Studien einschließlich aller verfügbaren Quellen in Abschnitt 4.3.1.4 auf. Alle durch die vorhergehenden Schritte identifizierten und in der Tabelle genannten Quellen der relevanten Studien sollen für die Bewertung dieser Studien herangezogen werden.

Folgende Informationen sind in der Tabelle darzulegen: Studienbezeichnung, Studienkategorie und verfügbare Quellen. Orientieren Sie sich dabei an der beispielhaften Angabe in der ersten Tabellenzeile. Hierbei sollen die Studien durch Zwischenzeilenüberschriften ggf. sinnvoll angeordnet werden, beispielsweise nach Therapieschema (Akut-/Langzeitstudien) und jeweils separat nach Art der Kontrolle (Placebo, zweckmäßige Vergleichstherapie, beides). Sollten Sie eine Strukturierung des Studienpools vornehmen, berücksichtigen Sie diese auch in den weiteren Tabellen in Modul 4.

Tabelle 4-5: Studienpool – RCT mit dem zu bewertenden Arzneimittel

Studie	Studienkategorie			verfügbare Quellen[a]		
	Studie zur Zulassung des zu bewertenden Arzneimittels (ja/nein)	gesponserte Studie[b] (ja/nein)	Studie Dritter (ja/nein)	Studienberichte (ja/nein [Zitat])	Register-einträge[c] (ja/nein [Zitat])	Publikation und sonstige Quellen[d] (ja/nein [Zitat])
ggf. Zwischenüberschrift zur Strukturierung des Studienpools						
placebokontrolliert						
<Studie 1>	ja	ja	nein	ja [5]	ja [6, 7]	ja [8]
aktivkontrolliert, zweckmäßige Vergleichstherapie(n)						

a: Bei Angabe „ja“ sind jeweils die Zitate der Quelle(n) (z. B. Publikationen, Studienberichte, Studienregister-einträge) mit anzugeben, und zwar als Verweis auf die in Abschnitt 4.6 genannte Referenzliste. Darüber hinaus ist darauf zu achten, dass alle Quellen, auf die in dieser Tabelle verwiesen wird, auch in Abschnitt 4.3.1.4 (Liste der eingeschlossenen Studien) aufgeführt werden.

b: Studie, für die der Unternehmer Sponsor war.

c: Zitat der Studienregistereinträge sowie, falls vorhanden, der in den Studienregistern aufgelisteten Berichte über Studiendesign und/oder -ergebnisse.

d: Sonstige Quellen: Dokumente aus der Suche auf der Internetseite des G-BA.

4.3.1.2 Charakteristika der in die Bewertung eingeschlossenen Studien – RCT mit dem zu bewertenden Arzneimittel

4.3.1.2.1 Studiendesign und Studienpopulationen

Beschreiben Sie das Studiendesign und die Studienpopulation der in die Bewertung eingeschlossenen Studien mindestens mit den Informationen in den folgenden Tabellen. Falls Teilpopulationen berücksichtigt werden, ist die Charakterisierung der Studienpopulation auch für diese Teilpopulation durchzuführen. Orientieren Sie sich dabei an der beispielhaften Angabe in der ersten Tabellenzeile. Geben Sie bei den Datenschnitten auch den Anlass des Datenschnittes an. Fügen Sie für jede Studie eine neue Zeile ein.

Weitere Informationen zu Studiendesign, Studienmethodik und Studienverlauf sind in Anhang 4-E zu hinterlegen.

Tabelle 4-6: Charakterisierung der eingeschlossenen Studien – RCT mit dem zu bewertenden Arzneimittel

Studie	**Studiendesign** <RCT, doppelblind/einfach, verblindet/offen, parallel/cross-over etc.>	**Population** <relevante Charakteristika, z. B. Schweregrad>	**Interventionen (Zahl der randomisierten Patienten)**	**Studiendauer/ Datenschnitte** <ggf. Run-in, Behandlung, Nachbeobachtung>	**Ort und Zeitraum der Durchführung**	**Primärer Endpunkt; patientenrelevante sekundäre Endpunkte**
<Studie 1>	RCT, doppelblind, parallel	Jugendliche und Erwachsene, leichtes bis mittelschweres Asthma	<Gruppe 1> (n= 354) <Gruppe 2> (n= 347)	Run-in: 2 Wochen Behandlung: 6 Monate 1. Datenschnitt: 1.7.2015 (z.B. geplante Interimsanalyse) 2. Datenschnitt: 1.1.2016 (z.B. Anforderung EMA, ungeplant)	Europa (Deutschland, Frankreich, Polen) 9/2003 – 12/2004	FEV1; Asthma-Symptome, Exazerbationen, gesundheitsbezogene Lebensqualität, unerwünschte Ereignisse

Tabelle 4-7: Charakterisierung der Interventionen – RCT mit dem zu bewertenden Arzneimittel

Studie	**<Gruppe 1>**	**<Gruppe 2>**	***ggf. weitere Spalten mit Behandlungscharakteristika*** ***z. B. Vorbehandlung, Behandlung in der Run-in-Phase etc.***
<Studie 1>	xxx 250 µg, 1 Inhalation bid + Placebo 2 Inhalationen bid	yyy 200 µg, 2 Inhalationen bid + Placebo 1 Inhalation bid	Vorbehandlung: zzz 1000 µg pro Tag, 4 Wochen vor Studienbeginn Bedarfsmedikation: aaa

Tabelle 4-8: Charakterisierung der Studienpopulationen – RCT mit dem zu bewertenden Arzneimittel

Studie **Gruppe**	**N**	**Alter (Jahre)**	**Geschlecht w/m (%)**	***ggf. weitere Spalten mit Populationscharakteristika*** ***z. B. Dauer der Erkrankung, Schweregrad, Therapieabbrecher, Studienabbrecher, weitere Basisdaten projektabhängig***
<Studie 1> <Gruppe 1> <Gruppe 2>				

Beschreiben Sie die Studien zusammenfassend. In der Beschreibung der Studien sollten Informationen zur Behandlungsdauer sowie zu geplanter und tatsächlicher Beobachtungsdauer enthalten sein. Sofern sich die Beobachtungsdauer zwischen den relevanten Endpunkten unterscheidet, sind diese unterschiedlichen Beobachtungsdauern endpunktbezogen anzugeben. Beschreiben Sie zudem, ob und aus welchem Anlass verschiedene Datenschnitte durchgeführt wurden oder noch geplant sind. Geben Sie dabei auch an, ob diese Datenschnitte jeweils vorab (d.h. im statistischen Analyseplan) geplant waren. In der Regel ist nur die Darstellung von a priori geplanten oder von Zulassungsbehörden geforderten Datenschnitten erforderlich. Machen Sie auch Angaben zur Übertragbarkeit der Studienergebnisse auf den deutschen Versorgungskontext.

Sollte es Unterschiede zwischen den Studien geben, weisen Sie in einem erläuternden Text darauf hin.

<< Angaben des pharmazeutischen Unternehmers >>

4.3.1.2.2 Verzerrungspotenzial auf Studienebene

Bewerten Sie das Verzerrungspotenzial der RCT auf Studienebene mithilfe des Bewertungsbogens in Anhang 4-F. Fassen Sie die Bewertung mit den Angaben in der folgenden Tabelle zusammen. Fügen Sie für jede Studie eine neue Zeile ein.

Dokumentieren Sie die Einschätzung für jede Studie mit einem Bewertungsbogen in Anhang 4-F.

Tabelle 4-9: Verzerrungspotenzial auf Studienebene – RCT mit dem zu bewertenden Arzneimittel

Studie	Adäquate Erzeugung der Randomisierungs-sequenz	Verdeckung der Gruppenzuteilung	Verblindung		Ergebnisunabhängige Berichterstattung	Keine sonstigen Aspekte	Verzerrungspotenzial auf Studienebene
			Patient	Behandelnde Personen			
<Studie 1>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein>	<hoch / niedrig>

Begründen Sie für jede Studie die abschließende Einschätzung.

<< Angaben des pharmazeutischen Unternehmers >>

4.3.1.3 Ergebnisse aus randomisierten kontrollierten Studien

Geben Sie in der folgenden Tabelle einen Überblick über die patientenrelevanten Endpunkte, auf denen Ihre Bewertung des medizinischen Nutzens und Zusatznutzens beruht. Geben Sie dabei an, welche dieser Endpunkte in den relevanten Studien jeweils untersucht wurden. Orientieren Sie sich dabei an der beispielhaften Angabe in der ersten Tabellenzeile. Fügen Sie für jede Studie eine neue Zeile ein.

Tabelle 4-10: Matrix der Endpunkte in den eingeschlossenen RCT mit dem zu bewertenden Arzneimittel

Studie	<Mortalität>	<Gesundheits-bezogene Lebensqualität>	<Endpunkt>	<Endpunkt>	<Endpunkt>
<Studie 1>	nein	ja	ja	ja	nein

4.3.1.3.1 <Endpunkt xxx> – RCT

Die Ergebnisdarstellung für jeden Endpunkt umfasst 3 Abschnitte. Zunächst soll für jede Studie das Verzerrungspotenzial auf Endpunktebene in einer Tabelle zusammengefasst werden. Dann sollen die Ergebnisse der einzelnen Studien zu dem Endpunkt tabellarisch dargestellt und in einem Text zusammenfassend beschrieben werden. Anschließend sollen die Ergebnisse, wenn möglich und sinnvoll, in einer Meta-Analyse zusammengefasst und beschrieben werden.

Die tabellarische Darstellung der Ergebnisse für den jeweiligen Endpunkt soll mindestens die folgenden Angaben enthalten:

– Ergebnisse der ITT-Analyse
– Zahl der Patienten, die in die Analyse eingegangen sind inkl. Angaben zur Häufigkeit von und zum Umgang mit nicht oder nicht vollständig beobachteten Patienten (bei Verlaufsbeobachtungen pro Messzeitpunkt)
– dem Endpunkt entsprechende Kennzahlen pro Behandlungsgruppe
– bei Verlaufsbeobachtungen Werte zu Studienbeginn und Studienende inklusive Standardabweichung
– bei dichotomen Endpunkten die Anzahlen und Anteile pro Gruppe sowie Angabe des relativen Risikos, des Odds Ratios und der absoluten Risikoreduktion
– entsprechende Maße bei weiteren Messniveaus
– Effektschätzer mit zugehörigem Standardfehler
– Angabe der verwendeten statistischen Methodik inklusive der Angabe der Faktoren, nach denen ggf. adjustiert wurde.

Unterschiedliche Beobachtungszeiten zwischen den Behandlungsgruppen sollen durch adäquate Analysen (z.B. Überlebenszeitanalysen) adressiert werden, und zwar für alle Endpunkte (einschließlich UE nach den nachfolgend genannten Kriterien), für die eine solche Analyse aufgrund deutlich unterschiedlicher Beobachtungszeiten erforderlich ist.

Bei Überlebenszeitanalysen soll die Kaplan-Meier-Kurve einschließlich Angaben zu den Patienten unter Risiko im Zeitverlauf (zu mehreren Zeitpunkten) abgebildet werden. Dabei ist für jeden Endpunkt, für den eine solche Analyse durchgeführt wird, eine separate Kaplan-Meier-Kurve darzustellen.

Zu mit Skalen erhobenen patientenberichteten Endpunkten (z.B. zur gesundheitsbezogenen Lebensqualität oder zu Symptomen) sind immer auch die Werte im Studienverlauf anzugeben, auch als grafische Darstellung, sowie eine Auswertung, die die über den Studienverlauf ermittelten Informationen vollständig berücksichtigt (z.B. als Symptomlast über die Zeit, geschätzt mittels MMRM-Analyse [falls aufgrund der Datenlage geeignet]). Die Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen soll nach folgendem Vorgehen erfolgen:

1. Falls in einer Studie Responderanalysen unter Verwendung einer MID präspezifiziert sind und das Responsekriterium mindestens 15 % der Skalenspannweite des verwendeten

Erhebungsinstruments entspricht, sind diese Responderanalysen für die Bewertung darzustellen.

2. Falls präspezifiziert Responsekriterien im Sinne einer MID unterhalb von 15 % der Skalenspannweite liegen, bestehen in diesen Fällen und solchen, in denen gar keine Responsekriterien präspezifiziert wurden, aber stattdessen Analysen kontinuierlicher Daten zur Verfügung stehen, verschiedene Möglichkeiten. Entweder können die Analysen der kontinuierlichen Daten dargestellt werden, für die Relevanzbewertung ist dabei auf ein allgemeines statistisches Maß in Form von standardisierten Mittelwertdifferenzen (SMDs, in Form von Hedges' g) zurückzugreifen. Dabei ist eine Irrelevanzschwelle von 0,2 zu verwenden: Liegt das zum Effektschätzer korrespondierende Konfidenzintervall vollständig oberhalb dieser Irrelevanzschwelle, wird davon ausgegangen, dass die Effektstärke nicht in einem sicher irrelevanten Bereich liegt. Dies soll gewährleisten, dass der Effekt hinreichend sicher mindestens als klein angesehen werden kann. Alternativ können post hoc spezifizierte Analysen mit einem Responsekriterium von genau 15 % der Skalenspannweite dargestellt werden.

Zu unerwünschten Ereignissen (UE) sind folgende Auswertungen vorzulegen:

1. Gesamtrate UE,

2. Gesamtrate schwerwiegender UE (SUE),

3. Gesamtrate der Abbrüche wegen UE,

4. Gesamtraten von UE differenziert nach Schweregrad, sofern dies in der/den relevante/n Studie/n erhoben wurde (z.B. gemäß CTCAE und/oder einer anderen etablierten bzw. validierten indikationsspezifischen Klassifikation) einschließlich einer Abgrenzung schwerer und nicht schwerer UE,

5. zu den unter 1, 2 und 4 genannten Kategorien (UE ohne weitere Differenzierung, SUE, UE differenziert nach Schweregrad) soll zusätzlich zu den Gesamtraten die Darstellung nach Organsystemen und Einzelereignissen (als System Organ Class [SOCs] und Preferred Terms [PT] nach MedDRA) jeweils nach folgenden Kriterien erfolgen:

- UE (unabhängig vom Schweregrad): Ereignisse, die bei mindestens 10% der Patienten in einem Studienarm aufgetreten sind

- Schwere UE (z.B. CTCAE-Grad ≥ 3) und SUE: Ereignisse, die bei mindestens 5% der Patienten in einem Studienarm aufgetreten sind

- zusätzlich für alle Ereignisse unabhängig vom Schweregrad: Ereignisse, die bei mindestens 10 Patienten UND bei mindestens 1 % der Patienten in einem Studienarm aufgetreten sind.

6. A priori definierte UE von besonderem Interesse [AESI]) sowie prädefinierte SOC-übergreifende UE-Auswertungen (z.B. als Standardised MedDRA Queries, SMQs) sollen unabhängig von der Ereignisrate dargestellt werden und zwar differenziert nach Schweregrad

(dargestellt als Gesamtrate und differenziert nach Schweregrad, nicht schwer, schwer, schwerwiegend).

7. zu Kategorie 3: Die Abbruchgründe auf SOC/PT-Ebene müssen vollständig, jedoch nur deskriptiv dargestellt werden.

Sofern bei der Erhebung unerwünschter Ereignisse erkrankungsbezogenen Ereignisse (z. B. Progression, Exazerbation) berücksichtigt werden (diese Ereignisse also in die UE-Erhebung eingehen), sollen für die Gesamtraten (UE, schwere UE und SUE) zusätzliche UE-Analysen durchgeführt werden, bei denen diese Ereignisse unberücksichtigt bleiben. Alle Auswertungen zu UE können auch in einem separaten Anhang des vorliegenden Modul 4 dargestellt werden. Dabei kann die Ausgabe der Statistik-Software unverändert verwendet werden, sofern diese alle notwendigen Angaben enthält. Eine Darstellung ausschließlich in Modul 5 ist nicht ausreichend. Davon unbenommen sind die Gesamtraten (UE, schwere UE, SUE und Abbrüche wegen UE), sowie die für die Gesamtaussage zum Zusatznutzen herangezogenen Ergebnisse im vorliegenden Abschnitt darzustellen.

Auswertungen zu den im Abschnitt 4.3.1.2.1 aufgeführten Datenschnitten sollen vollständig, d.h. für alle erhobenen relevanten Endpunkte, durchgeführt und vorgelegt werden. Das gilt auch dann wenn ein Datenschnitt ursprünglich nur zur Auswertung einzelner Endpunkte geplant war. Auf die Darstellung der Ergebnisse einzelner Endpunkte eines Datenschnitts bzw. eines gesamten Datenschnitts kann verzichtet werden, wenn hierdurch kein wesentlicher Informationsgewinn gegenüber einem anderen Datenschnitt zu erwarten ist (z. B. wenn die Nachbeobachtung zu einem Endpunkt bereits zum vorhergehenden Datenschnitt nahezu vollständig war oder ein Datenschnitt in unmittelbarer zeitlicher Nähe zu einem anderen Datenschnitt liegt).

Falls für die Auswertung eine andere Population als die ITT-Population herangezogen wird, soll diese benannt (z.B. Safety-Population) und definiert werden.

Sofern mehrere Studien vorliegen, sollen diese in einer Meta-Analyse zusammengefasst werden, wenn die Studien aus medizinischen (z. B. Patientengruppen) und methodischen (z. B. Studiendesign) Gründen ausreichend vergleichbar sind. Es ist jeweils zu begründen, warum eine Meta-Analyse durchgeführt wurde oder warum eine Meta-Analyse nicht durchgeführt wurde bzw. warum einzelne Studien ggf. nicht in die Meta-Analyse einbezogen wurden. Sofern die vorliegenden Studien für eine Meta-Analyse geeignet sind, sollen die Meta-Analysen als Forest-Plot dargestellt werden. Die Darstellung soll ausreichende Informationen zur Einschätzung der Heterogenität der Ergebnisse zwischen den Studien in Form von geeigneten statistischen Maßzahlen enthalten (siehe Abschnitt 4.2.5.3). Eine Gesamtanalyse aller Patienten aus mehreren Studien ohne Berücksichtigung der Studienzugehörigkeit (z. B. Gesamt-Vierfeldertafel per Addition der Einzel-Vierfeldertafeln) soll vermieden werden, da so die Heterogenität nicht eingeschätzt werden kann.

Beschreiben Sie die Operationalisierung des Endpunkts für jede Studie in der folgenden Tabelle. Fügen Sie für jede Studie eine neue Zeile ein.

Tabelle 4-11: Operationalisierung von <Endpunkt xxx>

Studie	Operationalisierung
<Studie 1>	

Bewerten Sie das Verzerrungspotenzial für den in diesem Abschnitt beschriebenen Endpunkt mithilfe des Bewertungsbogens in Anhang 4-F. Fassen Sie die Bewertung mit den Angaben in der folgenden Tabelle zusammen. Fügen Sie für jede Studie eine neue Zeile ein.

Dokumentieren Sie die Einschätzung für jede Studie mit einem Bewertungsbogen in Anhang 4-F.

Tabelle 4-12: Bewertung des Verzerrungspotenzials für <Endpunkt xxx> in RCT mit dem zu bewertenden Arzneimittel

Studie	Verzerrungspotenzial auf Studienebene	Verblindung Endpunkterheber	Adäquate Umsetzung des ITT-Prinzips	Ergebnisunabhängige Berichterstattung	Keine sonstigen Aspekte	Verzerrungspotenzial Endpunkt
<Studie 1>	<hoch / niedrig>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein>	<hoch / niedrig>

Begründen Sie für jede Studie die abschließende Einschätzung.

<< Angaben des pharmazeutischen Unternehmers >>

Stellen Sie die Ergebnisse für den Endpunkt xxx für jede einzelne Studie in tabellarischer Form dar. Fügen Sie für jede Studie eine neue Zeile ein. Beschreiben Sie die Ergebnisse zusammenfassend.

Tabelle 4-13: Ergebnisse für <Endpunkt xxx> aus RCT mit dem zu bewertenden Arzneimittel

Studie	Tabellarische Präsentation in geeigneter Form (Anforderungen siehe Erläuterung oben)
<Studie 1>	

<< Angaben des pharmazeutischen Unternehmers >>

Sofern die vorliegenden Studien bzw. Daten für eine Meta-Analyse medizinisch und methodisch geeignet sind, fassen Sie die Einzelergebnisse mithilfe von Meta-Analysen quantitativ zusammen und stellen Sie die Ergebnisse der Meta-Analysen (in der Regel als Forest-Plot) dar. Beschreiben Sie die Ergebnisse zusammenfassend. Begründen Sie, warum eine Meta-Analyse durchgeführt wurde bzw. warum eine Meta-Analyse nicht durchgeführt wurde bzw. warum einzelne Studien ggf. nicht in die Meta-Analyse einbezogen wurden. Machen Sie auch Angaben zur Übertragbarkeit der Studienergebnisse auf den deutschen Versorgungskontext.

<Abbildung Meta-Analyse>

Abbildung 2: Meta-Analyse für <Endpunkt xxx> aus RCT; <zu bewertendes Arzneimittel> versus <Vergleichstherapie>

<< Angaben des pharmazeutischen Unternehmers >>

Stellen Sie die in diesem Abschnitt beschriebenen Informationen für jeden weiteren Endpunkt aus RCT mit dem zu bewertenden Arzneimittel fortlaufend in einem eigenen Abschnitt dar.

4.3.1.3.2 Subgruppenanalysen – RCT

Für die Darstellung der Ergebnisse aus Subgruppenanalysen gelten die gleichen Anforderungen wie für die Darstellung von Ergebnissen aus Gesamtpopulationen in Abschnitt 4.3.1.3.1. [16]

Darüber hinaus sind folgende Kriterien zu berücksichtigen:

- Subgruppenanalysen sind nur für die Merkmale (z.B. Alter) durchzuführen, bei denen die resultierenden Subgruppen jeweils mindestens 10 Patienten umfassen.
- Subgruppenanalysen sind für binäre Ereignisse je Merkmal nur dann durchzuführen, wenn in einer der Subgruppen mindestens 10 Ereignisse aufgetreten sind.
- Für Überlebenszeitanalysen müssen Kaplan-Meier-Kurven zu den einzelnen Subgruppen nur für Subgruppenanalysen mit statistisch signifikantem Interaktionsterm ($p < 0,05$) dargestellt werden.

[16] unbesetzt

- Ergebnisse zu UE nach SOC und PT müssen nur dargestellt werden, wenn das jeweilige Ergebnis für die Gesamtpopulation statistisch signifikant ist. Zu a priori definierten Ereignissen (z.B. AESI, SMQs) sowie den UE-Gesamtraten (UE, schwere UE, SUE und Abbrüche wegen UE) müssen Subgruppenanalysen unabhängig vom Vorliegen statistischer Signifikanz in der Gesamtpopulation dargestellt werden.

- Bei Vorliegen mehrerer Studien und Durchführung von Metaanalysen zu diesen Studien gelten die zuvor genannten Kriterien für die jeweilige Metaanalyse, nicht für die Einzelstudien.

- Für Studien des pharmazeutischen Unternehmers sind entsprechende Analysen für alle benannten Effektmodifikatoren zu allen relevanten Endpunkten nach den zuvor genannten Kriterien vorzulegen und daher ggf. posthoc durchzuführen.

- Wird für die Nutzenbewertung nur die Teilpopulation einer Studie herangezogen (z.B. wegen Zulassungsbeschränkungen, aufgrund von durch den G-BA bestimmte Teilpopulationen), so gelten die genannten Kriterien für diese Teilpopulation, und die Subgruppenanalysen sind für die Teilpopulation und nicht für die Gesamtpopulation der Studie durchzuführen.

- Subgruppenanalysen, bei denen der Interaktionsterm nicht statistisch signifikant ist, können auch in einem separaten Anhang des vorliegenden Modul 4 dargestellt werden. Dabei kann die Ausgabe der Statistik-Software unverändert verwendet werden, sofern diese alle notwendigen Angaben enthält. Eine ausschließliche Darstellung in Modul 5 ist aber nicht ausreichend.

Beschreiben Sie die Ergebnisse von Subgruppenanalysen. Stellen Sie dabei zunächst tabellarisch dar, zu welchen der in Abschnitt 4.2.5.5 genannten Effektmodifikatoren Subgruppenanalysen zu den relevanten Endpunkten vorliegen, und ob diese a priori geplant und im Studienprotokoll festgelegt waren oder posthoc durchgeführt wurden.

Orientieren Sie sich an der beispielhaften Angabe in der ersten Tabellenzeile.

Tabelle 4 -14 Matrix der durchgeführten Subgruppenanalysen

Endpunkt Studie	**Alter**	**Geschlecht**	**<Effektmo-difikator-a>**	**<Effektmo-difikator-b>**	**<Effektmo-difikator-c>**	**<Effektmo-difikator-d>**
Gesamtmortalität						
<Studie 1>	●	●	●	○	○	○
<Studie 2>	●	●	○	n.d.	n.d.	n.d.
<Endpunkt 2>						
...						
●: A priori geplante Subgruppenanalyse. ○: Posthoc durchgeführte Subgruppenanalyse. n.d.: Subgruppenanalyse nicht durchgeführt.						

Stellen Sie anschließend in Tabelle 4-15 die Ergebnisse der Interaktionsterme für alle Subgruppenanalysen je Endpunkt in tabellarischer Form dar, und zwar für jede einzelne Studie separat. Kennzeichnen Sie dabei statistisch signifikante (p < 0,05) Interaktionsterme.

Tabelle 4-15: Ergebnis des Interaktionsterms der Subgruppenanalysen je Endpunkt für <Studie> und <Effektmodifikator>

Endpunkt Studie	**Alter**	**Geschlecht**	**<Effektmo-difikator-a>**	**<Effektmo-difikator-b>**	**<Effektmo-difikator-c>**	**<Effektmo-difikator-d>**
Gesamtmortalität						
<Studie 1>	p=0,345	p=0,321	**p=0,003**	**p=0,041**	p=0,981	p=0,212
<Studie 2>	p=0,634	p=0,212	**p<0,001**	k.A.	k.A.	k.A.
<Endpunkt 2>						
...						
k.A.: keine Angabe.						

Stellen Sie schließlich alle Subgruppenergebnisse dar.

Sofern eine Effektmodifikation für mehr als ein Subgruppenmerkmal vorliegt, kann eine Untersuchung auf eine Wechselwirkung höherer Ordnung sinnvoll sein. Dies gilt insbesondere dann, wenn diese Effektmodifikation konsistent über mehrere Endpunkte besteht. Zur Interpretation der Ergebnisse sollte dann für diese Endpunkte zusätzlich eine Subgruppenanalyse durchgeführt werden, die die Merkmale mit Effektmodifikation kombiniert. Beispiel: Für die Endpunkte Mortalität, gesundheitsbezogene Lebensqualität und schwere unerwünschte Ereignisse liegt sowohl für das Merkmal Geschlecht (mit den Ausprägungen „weiblich“ und „männlich“) als auch für das Merkmal Schweregrad (mit den Ausprägungen „niedrig“ und „hoch“) eine Effektmodifikation vor. Die zusätzliche Subgruppenanalyse erfolgt dann für die 3 genannten Endpunkte für das kombinierte Merkmal Geschlecht/Schweregrad mit den 4 Ausprägungen weiblich/niedrig, weiblich/hoch, männlich/niedrig und männlich/hoch.

Sofern die vorliegenden Studien bzw. Daten für eine Meta-Analyse medizinisch und methodisch geeignet sind, fassen Sie die Ergebnisse mithilfe einer Meta-Analyse quantitativ zusammen und stellen Sie die Ergebnisse der Meta-Analyse (als Forest-Plot) dar.

Beschreiben Sie die Ergebnisse zusammenfassend. Begründen Sie Ihr Vorgehen, wenn Sie keine Meta-Analyse durchführen bzw. wenn Sie nicht alle Studien in die Meta-Analyse einschließen.

<< Angaben des pharmazeutischen Unternehmers >>

4.3.1.4 Liste der eingeschlossenen Studien - RCT

Listen Sie alle für die Nutzenbewertung berücksichtigten Studien und Untersuchungen unter Angabe der im Dossier verwendeten Studienbezeichnung und der zugehörigen Quellen (z. B. Publikationen, Studienberichte, Studienregistereinträge).

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2 Weitere Unterlagen

4.3.2.1 Indirekte Vergleiche auf Basis randomisierter kontrollierter Studien

Hinweis: Die nachfolgenden Unterabschnitte sind nur dann auszufüllen, wenn indirekte Vergleiche als Nachweis für einen Zusatznutzen herangezogen werden sollen. Das ist dann möglich, wenn keine direkten Vergleichsstudien für das zu bewertende Arzneimittel gegenüber der zweckmäßigen Vergleichstherapie vorliegen oder diese keine ausreichenden Aussagen über den Zusatznutzen zulassen.

4.3.2.1.1 Ergebnis der Informationsbeschaffung – Studien für indirekte Vergleiche

Beschreiben Sie nachfolgend das Ergebnis der Informationsbeschaffung zu Studien für indirekte Vergleiche. ***Strukturieren Sie diesen Abschnitt analog Abschnitt 4.3.1.1 (Ergebnis der Informationsbeschaffung – RCT mit dem zu bewertenden Arzneimittel) und stellen Sie Informationen sowohl für das zu bewertende Arzneimittel als auch für die zweckmäßige Vergleichstherapie analog Abschnitt 4.3.1.1 zur Verfügung (einschließlich tabellarischer Darstellungen, Angabe eines Flussdiagramms etc.).*** *Benennen Sie sowohl für das zu bewertende Arzneimittel als auch für die zweckmäßige Vergleichstherapie*

- *Studien des pharmazeutischen Unternehmers*
- *Studien aus der bibliografischen Literaturrecherche*
- *Studien aus der Suche in Studienregistern/ Studienergebnisdatenbanken*
- *Studien aus der Suche auf der Internetseite des G-BA*
- *Resultierender Studienpool aus den einzelnen Suchschritten*

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.1.2 Charakteristika der Studien für indirekte Vergleiche

Charakterisieren Sie nachfolgend die Studien, die für indirekte Vergleiche identifiziert wurden und bewerten Sie darüber hinaus deren Ähnlichkeit. Begründen Sie darauf basierend den Ein- bzw. Ausschluss von Studien für die von Ihnen durchgeführten indirekten Vergleiche. Bewerten Sie das Verzerrungspotenzial der für indirekte Vergleiche herangezogenen Studien.

Strukturieren Sie diesen Abschnitt analog Abschnitt 4.3.1.2 und stellen Sie Informationen analog Abschnitt 4.3.1.2 zur Verfügung.

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.1.3 Ergebnisse aus indirekten Vergleichen

Geben Sie in der folgenden Tabelle einen Überblick über die patientenrelevanten Endpunkte, auf denen Ihre Bewertung des medizinischen Nutzens und Zusatznutzens aus indirekten Vergleichen beruht. Orientieren Sie sich dabei an der beispielhaften Angabe in der ersten Zeile. Fügen Sie für jede Studie eine neue Zeile ein.

Tabelle 4-16: Matrix der Endpunkte in den eingeschlossenen RCT für indirekte Vergleiche

Studie	**<Mortalität>**	**<Gesundheits-bezogene Lebensqualität>**	**<Endpunkt>**	**<Endpunkt>**	**<Endpunkt>**
<Studie 1>	nein	ja	ja	ja	nein

4.3.2.1.3.1 <Endpunkt xxx> – indirekte Vergleiche aus RCT

Für die indirekten Vergleiche soll zunächst für jeden Endpunkt eine Übersicht über die verfügbaren Vergleiche gegeben werden. Anschließend soll die Darstellung der Ergebnisse in drei Schritten erfolgen: 1) Bewertung des Verzerrungspotenzials auf Endpunktebene pro Studie, 2) tabellarische Darstellung der Ergebnisse der einzelnen Studien, 3) Darstellung des indirekten Vergleichs. **Für die Punkte 1 und 2 gelten die gleichen Anforderungen wie für die Darstellung der Ergebnisse der direkten Vergleiche in Abschnitt 4.3.1.3.1.**

Geben Sie für den im vorliegenden Abschnitt präsentierten Endpunkt einen Überblick über die in den Studien verfügbaren Vergleiche. Beispielhaft wäre folgende Darstellung denkbar:

Tabelle 4-17: Zusammenfassung der verfügbaren Vergleiche in den Studien, die für den indirekten Vergleich herangezogen wurden

Anzahl Studien	**Studie**	**Intervention**	**<Vergleichs-therapie 1>**	**<Vergleichs-therapie 2>**	**<Vergleichs-therapie 3>**
1	<Studie 1>	●		●	●
2	<Studie 2>	●		●	
	<Studie 3>	●		●	
1	<Studie 4>		●	●	●
etc.	etc.	etc.	etc.		

Stellen Sie zusätzlich die Netzwerkstruktur des indirekten Vergleichs grafisch dar.

<< Angaben des pharmazeutischen Unternehmers >>

Beschreiben Sie die Operationalisierung des Endpunkts für jede Studie in der folgenden Tabelle. Fügen Sie für jede Studie eine neue Zeile ein.

Tabelle 4-18: Operationalisierung von <Endpunkt xxx>

Studie	**Operationalisierung**
<Studie 1>	

Bewerten Sie das Verzerrungspotenzial für den in diesem Abschnitt beschriebenen Endpunkt mithilfe des Bewertungsbogens in Anhang 4-F. Fassen Sie die Bewertung mit den Angaben in der folgenden Tabelle zusammen. Fügen Sie für jede Studie eine neue Zeile ein.

Dokumentieren Sie die Einschätzung für jede Studie mit einem Bewertungsbogen in Anhang 4-F.

Tabelle 4-19: Bewertung des Verzerrungspotenzials für <Endpunkt xxx> in RCT für indirekte Vergleiche

Studie	**Verzerrungspotenzial auf Studienebene**	**Verblindung Endpunkterheber**	**Adäquate Umsetzung des ITT-Prinzips**	**Ergebnisunabhängige Berichterstattung**	**Keine sonstigen Aspekte**	**Verzerrungspotenzial Endpunkt**
<Studie 1>	<hoch / niedrig>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein>	<hoch / niedrig>

Begründen Sie für jede Studie die abschließende Einschätzung.

<< Angaben des pharmazeutischen Unternehmers >>

Stellen Sie die Ergebnisse für den Endpunkt xxx für jede einzelne Studie in tabellarischer Form dar. Fügen Sie für jede Studie eine neue Zeile ein. Beschreiben Sie die Ergebnisse zusammenfassend.

Tabelle 4-20: Ergebnisse für <Endpunkt xxx> aus RCT für indirekte Vergleiche

Studie	**Tabellarische Präsentation in geeigneter Form (Anforderungen siehe Erläuterung in Abschnitt 4.3.1.3.1)**
<Studie 1>	

<< Angaben des pharmazeutischen Unternehmers >>

Stellen Sie die Ergebnisse der indirekten Vergleiche in tabellarischer Form dar. Optional können die Ergebnisse zusätzlich auch grafisch illustriert werden. Orientieren Sie sich dabei an der üblichen Darstellung metaanalytischer Ergebnisse. Gliedern Sie die Ergebnisse nach folgenden Punkten:

- *Homogenität der Ergebnisse: Stellen Sie die Ergebnisse der paarweisen Meta-Analysen dar. Diskutieren Sie das Ausmaß sowie die Gründe für das Auftreten der Heterogenität für alle direkten paarweisen Vergleiche.*

- *Ergebnisse zu den Effekten: Stellen Sie die gepoolten Ergebnisse dar.*
- *Konsistenzprüfung: Stellen Sie die Ergebnisse der Konsistenzprüfung dar. Diskutieren Sie insbesondere inkonsistente Ergebnisse.*

Machen Sie darüber hinaus Angaben zur Übertragbarkeit der Studienergebnisse auf den deutschen Versorgungskontext.

<< Angaben des pharmazeutischen Unternehmers >>

Stellen Sie die in diesem Abschnitt beschriebenen Informationen für jeden weiteren Endpunkt für den ein indirekter Vergleich vorgenommen wird fortlaufend in einem eigenen Abschnitt dar.

4.3.2.1.3.2 Subgruppenanalysen – indirekte Vergleiche aus RCT

Beschreiben Sie nachfolgend die Ergebnisse von Subgruppenanalysen auf Basis indirekter Vergleiche aus RCT. ***Berücksichtigen Sie dabei die Anforderungen gemäß Abschnitt 4.3.1.3.2.***

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.1.4 Liste der eingeschlossenen Studien – indirekte Vergleiche aus RCT

Listen Sie alle für die Nutzenbewertung berücksichtigten Studien und Untersuchungen unter Angabe der im Dossier verwendeten Studienbezeichnung und der zugehörigen Quellen (z. B. Publikationen, Studienberichte, Studienregistereinträge).

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.2 Nicht randomisierte vergleichende Studien

Hinweis: Die nachfolgenden Unterabschnitte sind nur dann auszufüllen, wenn nicht randomisierte vergleichende Studien als Nachweis für einen Zusatznutzen herangezogen werden sollen.

4.3.2.2.1 Ergebnis der Informationsbeschaffung – nicht randomisierte vergleichende Studien

Beschreiben Sie nachfolgend das Ergebnis der Informationsbeschaffung zu nicht randomisierten vergleichenden Studien. ***Strukturieren Sie diesen Abschnitt analog Abschnitt 4.3.1.1 (Ergebnis der Informationsbeschaffung – RCT mit dem zu bewertenden Arzneimittel) und stellen Sie Informationen analog Abschnitt 4.3.1.1 zur Verfügung (einschließlich tabellarischer Darstellungen, Angabe eines Flussdiagramms etc.).*** *Benennen Sie*

- *Studien des pharmazeutischen Unternehmers*
- *Studien aus der bibliografischen Literaturrecherche*
- *Studien aus der Suche in Studienregistern/ Studienergebnisdatenbanken*
- *Studien aus der Suche auf der G-BA Internetseite*
- *Resultierender Studienpool aus den einzelnen Suchschritten*

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.2.2 Charakteristika der nicht randomisierten vergleichenden Studien

Charakterisieren Sie nachfolgend die nicht randomisierten vergleichenden Studien. ***Strukturieren Sie diesen Abschnitt analog Abschnitt 4.3.1.2 und stellen Sie Informationen analog Abschnitt 4.3.1.2 zur Verfügung.***

Beschreiben Sie die Verzerrungsaspekte der nicht randomisierten vergleichenden Studie auf Studienebene mithilfe des Bewertungsbogens in Anhang 4-F. Fassen Sie die Beschreibung mit den Angaben in der folgenden Tabelle zusammen. Fügen Sie für jede Studie eine neue Zeile ein.

Dokumentieren Sie die Einschätzung für jede Studie mit einem Bewertungsbogen in Anhang 4-F.

Tabelle 4-21: Verzerrungsaspekte auf Studienebene – nicht randomisierte vergleichende Interventionsstudien

Studie	**Zeitliche Parallelität der Gruppen**	**Vergleichbarkeit der Gruppen bzw. adäquate Berücksichtigung von prognostisch relevanten Faktoren**	**Verblindung**		**Ergebnisunabhängige Berichterstattung**	**Keine sonstigen Aspekte**
			Patient	**Behandelnde Personen**		
<Studie 1>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein>

Beschreiben Sie zusammenfassend die Bewertungsergebnisse zu Verzerrungsaspekten auf Studienebene.

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.2.3 Ergebnisse aus nicht randomisierten vergleichenden Studien

Geben Sie in der folgenden Tabelle einen Überblick über die patientenrelevanten Endpunkte, auf denen Ihre Bewertung des medizinischen Nutzens und Zusatznutzens aus nicht randomisierten vergleichenden Studien beruht. Orientieren Sie sich dabei an der beispielhaften Angabe in der ersten Zeile. Fügen Sie für jede Studie eine neue Zeile ein.

Tabelle 4-22: Matrix der Endpunkte in den eingeschlossenen nicht randomisierten vergleichenden Studien

Studie	**<Mortalität>**	**<Gesundheits-bezogene Lebensqualität>**	**<Endpunkt>**	**<Endpunkt>**	**<Endpunkt>**
<Studie 1>	nein	ja	ja	ja	nein

4.3.2.2.3.1 <Endpunkt xxx> – nicht randomisierte vergleichende Studien

Beschreiben Sie die Operationalisierung des Endpunkts für jede Studie in der folgenden Tabelle. Fügen Sie für jede Studie eine neue Zeile ein.

Tabelle 4-23: Operationalisierung von <Endpunkt xxx>

Studie	**Operationalisierung**
<Studie 1>	

Beschreiben Sie die Verzerrungsaspekte für den in diesem Abschnitt beschriebenen Endpunkt mithilfe des Bewertungsbogens in Anhang 4-F. Fassen Sie die Bewertung mit den Angaben in der folgenden Tabelle zusammen. Fügen Sie für jede Studie eine neue Zeile ein.

Dokumentieren Sie die Einschätzung für jede Studie mit einem Bewertungsbogen in Anhang 4-F.

Tabelle 4-24: Verzerrungsaspekte für <Endpunkt xxx> – nicht randomisierte vergleichende Studien

Studie	Verblindung Endpunkterheber	Adäquate Umsetzung des ITT-Prinzips	Ergebnisunabhängige Berichterstattung	Keine sonstigen Aspekte
<Studie 1>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein / unklar>	<ja / nein>

Beschreiben Sie zusammenfassend die Bewertungsergebnisse zu Verzerrungsaspekten auf Endpunktebene.

<< Angaben des pharmazeutischen Unternehmers >>

Stellen Sie die Ergebnisse der nicht randomisierten vergleichenden Studien gemäß den Anforderungen des TREND- bzw. des STROBE-Statements dar. Machen Sie dabei auch Angaben zur Übertragbarkeit der Studienergebnisse auf den deutschen Versorgungskontext.

<< Angaben des pharmazeutischen Unternehmers >>

Stellen Sie die in diesem Abschnitt beschriebenen Informationen für jeden weiteren Endpunkt aus nicht randomisierten vergleichenden Studien fortlaufend in einem eigenen Abschnitt dar.

4.3.2.2.3.2 Subgruppenanalysen – nicht randomisierte vergleichende Studien

Beschreiben Sie nachfolgend die Ergebnisse von Subgruppenanalysen aus nicht randomisierten vergleichenden Studien. ***Berücksichtigen Sie dabei die Anforderungen gemäß Abschnitt 4.3.1.3.2.***

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.2.4 Liste der eingeschlossenen Studien – nicht randomisierte vergleichende Studien

Listen Sie alle für die Nutzenbewertung berücksichtigten Studien und Untersuchungen unter Angabe der im Dossier verwendeten Studienbezeichnung und der zugehörigen Quellen (z. B. Publikationen, Studienberichte, Studienregistereinträge).

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.3 Weitere Untersuchungen

Hinweis: Die nachfolgenden Unterabschnitte sind nur dann auszufüllen, wenn über die in den Abschnitten 4.3.1, 4.3.2.1 und 4.3.2.2 genannten Studien hinausgehende Untersuchungen als Nachweis für einen Zusatznutzen herangezogen werden sollen.

4.3.2.3.1 Ergebnis der Informationsbeschaffung – weitere Untersuchungen

Beschreiben Sie nachfolgend das Ergebnis der Informationsbeschaffung nach Untersuchungen, die nicht in den Abschnitten 4.3.1, 4.3.2.1 und 4.3.2.2 aufgeführt sind. ***Strukturieren Sie diesen Abschnitt analog Abschnitt 4.3.1.1 (Ergebnis der Informationsbeschaffung – RCT mit dem zu bewertenden Arzneimittel) und stellen Sie Informationen sowohl für das zu bewertende Arzneimittel als auch für die zweckmäßige Vergleichstherapie analog Abschnitt 4.3.1.1 zur Verfügung (einschließlich tabellarischer Darstellungen, Angabe eines Flussdiagramms etc.).*** *Benennen Sie für das zu bewertende Arzneimittel als auch für die zweckmäßige Vergleichstherapie*

- *Studien des pharmazeutischen Unternehmers*
- *Studien aus der bibliografischen Literaturrecherche*
- *Studien aus der Suche in Studienregistern/ Studienergebnisdatenbanken*
- *Studien aus der Suche auf der G-BA Internetseite*
- *Resultierender Studienpool aus den einzelnen Suchschritten*

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.3.2 Charakteristika der weiteren Untersuchungen

Charakterisieren Sie nachfolgend die weiteren Untersuchungen und bewerten Sie deren Verzerrungsaspekte.

Ergebnisse nicht randomisierter Studien, die keine kontrollierten Interventionsstudien sind, gelten aufgrund ihres Studiendesigns generell als potenziell hoch verzerrt. Trifft das auf die von Ihnen vorgelegten Studien nicht zu, begründen Sie Ihre Einschätzung.

Strukturieren Sie diesen Abschnitt analog Abschnitt 4.3.1.2 und stellen Sie Informationen analog Abschnitt 4.3.1.2 zur Verfügung.

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.3.3 Ergebnisse aus weiteren Untersuchungen

Geben Sie in der folgenden Tabelle einen Überblick über die patientenrelevanten Endpunkte, auf denen Ihre Bewertung des medizinischen Nutzens und Zusatznutzens aus weiteren Untersuchungen beruht. Orientieren Sie sich dabei an der beispielhaften Angabe in der ersten Zeile. Fügen Sie für jede Studie eine neue Zeile ein.

Tabelle 4-25: Matrix der Endpunkte in den eingeschlossenen weiteren Untersuchungen

Studie	**<Mortalität>**	**<Gesundheits-bezogene Lebensqualität>**	**<Endpunkt>**	**<Endpunkt>**	**<Endpunkt>**
<Studie 1>	nein	ja	ja	ja	nein

4.3.2.3.3.1 <Endpunkt xxx> – weitere Untersuchungen

Beschreiben Sie die Operationalisierung des Endpunkts für jede Studie in der folgenden Tabelle. Fügen Sie für jede Studie eine neue Zeile ein.

Tabelle 4-26: Operationalisierung von <Endpunkt xxx> – weitere Untersuchungen

Studie	**Operationalisierung**
<Studie 1>	

Bewerten Sie die Verzerrungsaspekte für den in diesem Abschnitt beschriebenen Endpunkt. Ergebnisse nicht randomisierter Studien, die keine kontrollierten Interventionsstudien sind,

gelten aufgrund ihres Studiendesigns generell als potenziell hoch verzerrt. Trifft das auf die von Ihnen vorgelegten Studien nicht zu, begründen Sie Ihre Einschätzung.

<< Angaben des pharmazeutischen Unternehmers >>

Stellen Sie die Ergebnisse der weiteren Untersuchungen gemäß den jeweils gültigen Standards für die Berichterstattung dar. Begründen Sie dabei die Auswahl des Standards für die Berichterstattung. Machen Sie darüber hinaus Angaben zur Übertragbarkeit der Studienergebnisse auf den deutschen Versorgungskontext.

<< Angaben des pharmazeutischen Unternehmers >>

Stellen Sie die in diesem Abschnitt beschriebenen Informationen für jeden weiteren Endpunkt aus weiteren Untersuchungen fortlaufend in einem eigenen Abschnitt dar.

4.3.2.3.3.2 Subgruppenanalysen – weitere Untersuchungen

Beschreiben Sie nachfolgend die Ergebnisse von Subgruppenanalysen aus weiteren Untersuchungen. ***Berücksichtigen Sie dabei die Anforderungen gemäß Abschnitt 4.3.1.3.2.***

<< Angaben des pharmazeutischen Unternehmers >>

4.3.2.3.4 Liste der eingeschlossenen Studien – weitere Untersuchungen

Listen Sie alle für die Nutzenbewertung berücksichtigten Studien und Untersuchungen unter Angabe der im Dossier verwendeten Studienbezeichnung und der zugehörigen Quellen (z. B. Publikationen, Studienberichte, Studienregistereinträge).

<< Angaben des pharmazeutischen Unternehmers >>

4.4 Abschließende Bewertung der Unterlagen zum Nachweis des Zusatznutzens

4.4.1 Beurteilung der Aussagekraft der Nachweise

Legen Sie für alle im Dossier eingereichten Unterlagen die Evidenzstufe dar. Beschreiben Sie zusammenfassend auf Basis der in den Abschnitten 4.3.1 und 4.3.2 präsentierten Ergebnisse die Aussagekraft der Nachweise für einen Zusatznutzen unter Berücksichtigung der Studienqualität, der Validität der herangezogenen Endpunkte sowie der Evidenzstufe.

<< Angaben des pharmazeutischen Unternehmers >>

4.4.2 Beschreibung des Zusatznutzens einschließlich dessen Wahrscheinlichkeit und Ausmaß

Führen Sie die in den Abschnitten 4.3.1 und 4.3.2 beschriebenen Ergebnisse zum Zusatznutzen auf Ebene einzelner Endpunkte zusammen und leiten Sie ab, ob sich aus der Zusammenschau der Ergebnisse zu den einzelnen Endpunkten insgesamt ein Zusatznutzen des zu bewertenden Arzneimittels im Vergleich zur zweckmäßigen Vergleichstherapie ergibt. Berücksichtigen Sie dabei auch die Angaben zur Übertragbarkeit der Studienergebnisse auf den deutschen Versorgungskontext. Liegt ein Zusatznutzen vor, beschreiben Sie worin der Zusatznutzen besteht.

Stellen Sie die Wahrscheinlichkeit des Zusatznutzens dar, d. h., beschreiben und begründen Sie unter Berücksichtigung der in Abschnitt 4.4.1 dargelegten Aussagekraft der Nachweise die Ergebnissicherheit der Aussage zum Zusatznutzen.

Beschreiben Sie außerdem das Ausmaß des Zusatznutzens unter Verwendung folgender Kategorisierung (in der Definition gemäß AM-NutzenV):

- *erheblicher Zusatznutzen*
- *beträchtlicher Zusatznutzen*
- *geringer Zusatznutzen*
- *nicht quantifizierbarer Zusatznutzen*
- *kein Zusatznutzen belegbar*
- *der Nutzen des zu bewertenden Arzneimittels ist geringer als der Nutzen der zweckmäßigen Vergleichstherapie*

Berücksichtigen Sie bei den Aussagen zum Zusatznutzen ggf. nachgewiesene Unterschiede zwischen verschiedenen Patientengruppen.

<< Angaben des pharmazeutischen Unternehmers >>

4.4.3 Angabe der Patientengruppen, für die ein therapeutisch bedeutsamer Zusatznutzen besteht

Geben Sie auf Basis der in den Abschnitten 4.3.1 und 4.3.2 beschriebenen Ergebnisse und unter Berücksichtigung des in Abschnitt 4.4.2 dargelegten Zusatznutzens sowie dessen Wahrscheinlichkeit und Ausmaß in der nachfolgenden Tabelle an, für welche Patientengruppen ein therapeutisch bedeutsamer Zusatznutzen besteht. Benennen Sie das Ausmaß des Zusatznutzens in Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen. Fügen Sie für jede Patientengruppe mit therapeutisch bedeutsamem Zusatznutzen eine neue Zeile ein.

Tabelle 4-27: Patientengruppen, für die ein therapeutisch bedeutsamer Zusatznutzen besteht, einschließlich Ausmaß des Zusatznutzens

Bezeichnung der Patientengruppen	Ausmaß des Zusatznutzens

4.5 Begründung für die Vorlage weiterer Unterlagen und Surrogatendpunkte

4.5.1 Begründung für die Vorlage indirekter Vergleiche

Sofern mit dem Dossier indirekte Vergleiche (Abschnitt 4.3.2.1) eingereicht wurden, begründen Sie dies. Begründen Sie dabei auch, warum sich die ausgewählten Studien jeweils für einen indirekten Vergleich gegenüber dem zu bewertenden Arzneimittel und damit für den Nachweis eines Zusatznutzens durch indirekten Vergleich eignen.

<< Angaben des pharmazeutischen Unternehmers >>

4.5.2 Begründung für die Vorlage nicht randomisierter vergleichender Studien und weiterer Untersuchungen

Sofern mit dem Dossier nicht randomisierte vergleichende Studien (Abschnitt 4.3.2.2) oder weitere Untersuchungen (Abschnitt 4.3.2.3) eingereicht wurden, nennen Sie die Gründe, nach denen es unmöglich oder unangemessen ist, zu den in diesen Studien bzw. Untersuchungen behandelten Fragestellungen Studien höchster Evidenzstufe (randomisierte klinische Studien) durchzuführen oder zu fordern.

<< Angaben des pharmazeutischen Unternehmers >>

4.5.3 Begründung für die Bewertung auf Grundlage der verfügbaren Evidenz, da valide Daten zu patientenrelevanten Endpunkten noch nicht vorliegen

Falls aus Ihrer Sicht valide Daten zu patientenrelevanten Endpunkten zum Zeitpunkt der Bewertung noch nicht vorliegen können, begründen Sie dies.

<< Angaben des pharmazeutischen Unternehmers >>

4.5.4 Verwendung von Surrogatendpunkten

Die Verwendung von Surrogatendpunkten bedarf einer Begründung (siehe Abschnitt 4.5.3). Zusätzlich soll dargelegt werden, ob und warum die verwendeten Surrogatendpunkte im

betrachteten Kontext valide Surrogatendpunkte darstellen bzw. Aussagen zu patientenrelevanten Endpunkten zulassen.

Eine Validierung von Surrogatendpunkten bedarf in der Regel einer Meta-Analyse von Studien, in denen sowohl Effekte auf den Surrogatendpunkt als auch Effekte auf den interessierenden patientenrelevanten Endpunkt untersucht wurden (Burzykowski 2005[17], Molenberghs 2010[18]). Diese Studien müssen bei Patientenkollektiven und Interventionen durchgeführt worden sein, die Aussagen für das dem vorliegenden Antrag zugrundeliegende Anwendungsgebiet und das zu bewertende Arzneimittel sowie die Vergleichstherapie erlauben.

Eine Möglichkeit der Verwendung von Surrogatendpunkten ohne abschließende Validierung stellt die Anwendung des Konzepts eines sogenannten Surrogate-Threshold-Effekts (STE) (Burzykowski 2006[19]) dar. Daneben besteht die Möglichkeit einer Surrogatvalidierung in der quantitativen Betrachtung geeigneter Korrelationsmaße von Surrogatendpunkt und interessierendem patientenrelevanten Endpunkt („individuelle Ebene“) sowie von Effekten auf den Surrogatendpunkt und Effekten auf den interessierenden patientenrelevanten Endpunkt („Studienebene“). Dabei ist dann zu zeigen, dass die unteren Grenzen der entsprechenden 95%- Konfidenzintervalle für solche Korrelationsmaße ausreichend hoch sind. Die Anwendung alternativer Methoden zur Surrogatvalidierung (siehe Weir 2006[20]) soll ausreichend begründet werden, insbesondere dann, wenn als Datengrundlage nur eine einzige Studie verwendet werden soll.

Berichten Sie zu den Studien zur Validierung oder zur Begründung für die Verwendung von Surrogatendpunkten mindestens folgende Informationen:

– Patientenpopulation
– Intervention
– Kontrolle
– Datenherkunft
– verwendete Methodik
– entsprechende Ergebnisse (abhängig von der Methode)
– Untersuchungen zur Robustheit
– ggf. Untersuchungen zur Übertragbarkeit

Sofern Sie im Dossier Ergebnisse zu Surrogatendpunkten eingereicht haben, benennen Sie die Gründe für die Verwendung von Surrogatendpunkten. Beschreiben Sie, ob und warum die

[17] Burzykowski T (Ed.): The evaluation of surrogate endpoints. New York: Springer; 2005.
[18] Molenberghs G, Burzykowski T, Alonso A, Assam P, Tilahun A, Buyse M: A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials. Stat Methods Med Res 2010; 19(3): 205-236.
[19] Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharm Stat 2006; 5(3): 173-186.
[20] Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. Stat Med 2006; 25(2): 183-203.

verwendeten Surrogatendpunkte im betrachteten Kontext valide Surrogatendpunkte darstellen bzw. Aussagen zu patientenrelevanten Endpunkten zulassen.

<< Angaben des pharmazeutischen Unternehmers >>

4.6 Referenzliste

Listen Sie nachfolgend alle Quellen (z. B. Publikationen, Studienberichte, Studienregistereinträge), die Sie im vorliegenden Dokument angegeben haben (als fortlaufend nummerierte Liste). Verwenden Sie hierzu einen allgemein gebräuchlichen Zitierstil (z. B. Vancouver oder Harvard). Geben Sie bei Fachinformationen immer den Stand des Dokuments an.

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-A: Suchstrategien – bibliografische Literaturrecherche

Geben Sie nachfolgend die Suchstrategien für die bibliografische(n) Literaturrecherche(n) an, und zwar getrennt für die einzelnen Recherchen (Suche nach RCT mit dem zu bewertenden Arzneimittel, Suche nach RCT für indirekte Vergleiche etc.). Für jede durchsuchte Datenbank ist die verwendete Strategie separat darzustellen. Geben Sie dabei zunächst jeweils den Namen der durchsuchten Datenbank (z. B. EMBASE), die verwendete Suchoberfläche (z. B. DIMDI, Ovid etc.), das Datum der Suche, das Zeitsegment (z. B.: „1980 to 2010 week 50“) und die gegebenenfalls verwendeten Suchfilter (mit Angabe einer Quelle) an. Listen Sie danach die Suchstrategie einschließlich der resultierenden Trefferzahlen auf. Orientieren Sie sich bei Ihren Angaben an dem nachfolgenden Beispiel (eine umfassende Suche soll Freitextbegriffe und Schlagwörter enthalten):

Datenbankname	EMBASE	
Suchoberfläche	Ovid	
Datum der Suche	07.11.2016	
Zeitsegment	1974 to 2016 November 04	
Suchfilter	Filter für randomisierte kontrollierte Studien nach Wong 2006 [Quelle[21]] – Strategy minimizing difference between sensitivity and specificity	
#	**Suchbegriffe**	**Ergebnis**
1	Diabetes Mellitus/	552986
2	Non Insulin Dependent Diabetes Mellitus/	195234
3	(diabet* or niddm or t2dm).ab,ti.	714228
4	or/1-3	847068
5	linagliptin*.mp.	1562
6	(random* or double-blind*).tw.	1193849
7	placebo*.mp.	388057
8	or/6-7	1382838
9	and/4,5,8	633

[21] Das Zitat zu dem hier beispielhaft angegebenen Suchfilter lautet wie folgt: Wong SSL, Wilczynski NL, Haynes RB. Comparison of top-performing search strategies for detecting clinically sound treatment studies and systematic reviews in MEDLINE and EMBASE. J Med Libr Assoc 2006; 94(4): 451-455. Hinweis: Für die Suche in der Cochrane-Datenbank „Cochrane Central Register of Controlled Trials (Clinical Trials)“ sollte kein Studienfilter verwendet werden.

Anhang 4-A1: Suche nach RCT mit dem zu bewertenden Arzneimittel

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-A2: Suche nach RCT für indirekte Vergleiche

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-A3: Suche nach nicht randomisierten vergleichenden Studien

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-A4: Suche nach weiteren Untersuchungen

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-B: Suchstrategien – Suche in Studienregistern/ Studienergebnisdatenbanken

Geben Sie nachfolgend die Suchstrategien für die Suche(n) in Studienregistern/ Studienergebnisdatenbanken an. Machen Sie die Angaben getrennt für die einzelnen Recherchen (Suche nach RCT mit dem zu bewertenden Arzneimittel, Suche nach RCT für indirekte Vergleiche etc.) wie unten angegeben. Für jede/s durchsuchte Studienregister/ Studienergebnisdatenbank ist eine separate Strategie darzustellen. Geben Sie dabei jeweils den Namen des durchsuchten Studienregisters/ Studienergebnisdatenbank (z. B. clinicaltrials.gov), die Internetadresse, unter der das/die Studienregister/ Studienergebnisdatenbank erreichbar ist (z. B. http://www.clinicaltrials.gov), das Datum der Suche, die verwendete Suchstrategie und die resultierenden Treffer an. Orientieren Sie sich bei Ihren Angaben an dem nachfolgenden Beispiel:

Studienregister/ Studienergebnisdatenbank	International Clinical Trials Registry Platform Search Portal
Internetadresse	http://apps.who.int/trialsearch/
Datum der Suche	07.11.2016
Eingabeoberfläche	Standard Search
Suchstrategie	linagliptin OR BI 1356
Treffer	169

Anhang 4-B1: Suche nach RCT mit dem zu bewertenden Arzneimittel

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-B2: Suche nach RCT für indirekte Vergleiche

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-B3: Suche nach nicht randomisierten vergleichenden Studien

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-B4: Suche nach weiteren Untersuchungen

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-C: Liste der im Volltext gesichteten und ausgeschlossenen Dokumente mit Ausschlussgrund (bibliografische Literaturrecherche)

Listen Sie nachfolgend die im Volltext gesichteten und ausgeschlossenen Dokumente aus der /den bibliografischen Literaturrecherche(n) auf. Machen Sie die Angaben getrennt für die einzelnen Recherchen (Suche nach RCT mit dem zu bewertenden Arzneimittel, Suche nach RCT für indirekte Vergleiche etc.) wie unten angegeben. Verwenden Sie hierzu einen allgemein gebräuchlichen Zitierstil (z. B. Vancouver oder Harvard) und nummerieren Sie die Zitate fortlaufend. Geben Sie jeweils einen Ausschlussgrund an und beziehen Sie sich dabei auf die im Abschnitt 4.2.2 genannten Ein- und Ausschlusskriterien.

Anhang 4-C1: Suche nach RCT mit dem zu bewertenden Arzneimittel

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-C2: Suche nach RCT für indirekte Vergleiche

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-C3: Suche nach nicht randomisierten vergleichenden Studien

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-C4: Suche nach weiteren Untersuchungen

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-D: Liste der ausgeschlossenen Studien mit Ausschlussgrund (Suche in Studienregistern/ Studienergebnisdatenbanken)

Listen Sie nachfolgend die durch die Studienregistersuche(n)/ Studienergebnisdatenbanksuche(n) identifizierten, aber ausgeschlossenen Registereinträgen auf. Machen Sie die Angaben getrennt für die einzelnen Recherchen (Suche nach RCT mit dem zu bewertenden Arzneimittel, Suche nach RCT für indirekte Vergleiche etc.) wie unten angegeben. Verwenden Sie hierzu einen allgemein gebräuchlichen Zitierstil (z. B. Vancouver oder Harvard) und nummerieren Sie die Zitate fortlaufend. Geben Sie jeweils einen Ausschlussgrund an und beziehen Sie sich dabei auf die im Abschnitt 4.2.2 genannten Ein- und Ausschlusskriterien.

Anhang 4-D1: Suche nach RCT mit dem zu bewertenden Arzneimittel

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-D2: Suche nach RCT für indirekte Vergleiche

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-D3: Suche nach nicht randomisierten vergleichenden Studien

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-D4: Suche nach weiteren Untersuchungen

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-E: Methodik der eingeschlossenen Studien – RCT

Beschreiben Sie nachfolgend die Methodik jeder eingeschlossenen, in Abschnitt 4.3.1.1.5 genannten Studie. Erstellen Sie hierfür je Studie eine separate Version der nachfolgend dargestellten Tabelle 4-28 inklusive eines Flow-Charts für den Patientenfluss.

Sollten Sie im Dossier indirekte Vergleiche präsentieren, beschreiben Sie ebenfalls die Methodik jeder zusätzlich in den indirekten Vergleich eingeschlossenen Studie (Abschnitt 4.3.2.1). Erstellen Sie hierfür je Studie eine separate Version der nachfolgend dargestellten Tabelle 4-28 inklusive eines Flow-Charts für den Patientenfluss.

Tabelle 4-28 (Anhang): Studiendesign und -methodik für Studie <Studienbezeichnung>

Item[a]	Charakteristikum	Studieninformation
Studienziel		
2 b	Genaue Ziele, Fragestellung und Hypothesen	
Methoden		
3	Studiendesign	
3a	Beschreibung des Studiendesigns (z. B. parallel, faktoriell) inklusive Zuteilungsverhältnis	
3b	Relevante Änderungen der Methodik nach Studienbeginn (z. B. Ein-/Ausschlusskriterien), mit Begründung	
4	Probanden / Patienten	
4a	Ein-/Ausschlusskriterien der Probanden / Patienten	
4b	Studienorganisation und Ort der Studiendurchführung	
5	Interventionen Präzise Angaben zu den geplanten Interventionen jeder Gruppe und zur Administration etc.	
6	Zielkriterien	
6a	Klar definierte primäre und sekundäre Zielkriterien, Erhebungszeitpunkte, ggf. alle zur Optimierung der Ergebnisqualität verwendeten Erhebungsmethoden (z. B. Mehrfachbeobachtungen, Training der Prüfer) und ggf. Angaben zur Validierung von Erhebungsinstrumenten	
6b	Änderungen der Zielkriterien nach Studienbeginn, mit Begründung	
7	Fallzahl	
7a	Wie wurden die Fallzahlen bestimmt?	
7b	Falls notwendig, Beschreibung von Zwischenanalysen und Kriterien für einen vorzeitigen Studienabbruch	
8	Randomisierung, Erzeugung der Behandlungsfolge	
8a	Methode zur Generierung der zufälligen Zuteilung	
8b	Einzelheiten (z. B. Blockrandomisierung, Stratifizierung)	
9	Randomisierung, Geheimhaltung der Behandlungsfolge (allocation concealment) Durchführung der Zuteilung (z. B. nummerierte Behälter; zentrale Randomisierung per Fax / Telefon), Angabe, ob Geheimhaltung bis zur Zuteilung gewährleistet war	
10	Randomisierung, Durchführung Wer hat die Randomisierungsliste erstellt, wer nahm die Probanden/Patienten in die Studie auf und wer teilte die Probanden/Patienten den Gruppen zu?	

Item[a]	Charakteristikum	Studieninformation
11	Verblindung	
11a	Waren a) die Probanden / Patienten und / oder b) diejenigen, die die Intervention / Behandlung durchführten, und / oder c) diejenigen, die die Zielgrößen beurteilten, verblindet oder nicht verblindet, wie wurde die Verblindung vorgenommen?	
11b	Falls relevant, Beschreibung der Ähnlichkeit von Interventionen	
12	Statistische Methoden	
12a	Statistische Methoden zur Bewertung der primären und sekundären Zielkriterien	
12b	Weitere Analysen, wie z. B. Subgruppenanalysen und adjustierte Analysen	
Resultate		
13	Patientenfluss (inklusive Flow-Chart zur Veranschaulichung im Anschluss an die Tabelle)	
13a	Anzahl der Studienteilnehmer für jede durch Randomisierung gebildete Behandlungsgruppe, die a) randomisiert wurden, b) tatsächlich die geplante Behandlung/Intervention erhalten haben, c) in der Analyse des primären Zielkriteriums berücksichtigt wurden	
13b	Für jede Gruppe: Beschreibung von verlorenen und ausgeschlossenen Patienten nach Randomisierung mit Angabe von Gründen	
14	Aufnahme / Rekrutierung	
14a	Nähere Angaben über den Zeitraum der Studienaufnahme der Probanden / Patienten und der Nachbeobachtung	
14b	Informationen, warum die Studie endete oder beendet wurde	
a: nach CONSORT 2010.		

Stellen Sie für jede Studie den Patientenfluss in einem Flow-Chart gemäß CONSORT dar.

<< Angaben des pharmazeutischen Unternehmers >>

Anhang 4-F: Bewertungsbögen zur Einschätzung von Verzerrungsaspekten

Der nachfolgend dargestellte Bewertungsbogen dient der Dokumentation der Einstufung des Potenzials der Ergebnisse für Verzerrungen (Bias). Für jede Studie soll aus diesem Bogen nachvollziehbar hervorgehen, inwieweit die Ergebnisse für die einzelnen Endpunkte als möglicherweise verzerrt bewertet wurden, was die Gründe für die Bewertung waren und welche Informationen aus den Quellen dafür Berücksichtigung fanden.

Der Bogen gliedert sich in zwei Teile:

- Verzerrungsaspekte auf Studienebene. In diesem Teil sind die endpunktübergreifenden Kriterien aufgelistet.

- Verzerrungsaspekte auf Endpunktebene. In diesem Teil sind die Kriterien aufgelistet, die für jeden Endpunkt separat zu prüfen sind.

Für jedes Kriterium sind unter „Angaben zum Kriterium“ alle relevanten Angaben aus den Quellen zur Bewertung einzutragen (Stichworte reichen ggf., auf sehr umfangreiche Informationen in den Quellen kann verwiesen werden).

Grundsätzlich sollen die Bögen studienbezogen ausgefüllt werden. Wenn mehrere Quellen zu einer Studie vorhanden sind, müssen die herangezogenen Quellen in der folgenden Tabelle genannt und jeweils mit Kürzeln (z. B. A, B, C …) versehen werden. Quellenspezifische Angaben im weiteren Verlauf sind mit dem jeweiligen Kürzel zu kennzeichnen.

Hinweis: Der nachfolgend dargestellte Bewertungsbogen ist die Blankoversion des Bogens. Dieser Blankobogen ist für jede Studie heranzuziehen. Im Anschluss daran ist ein Bewertungsbogen inklusive Ausfüllhinweisen abgebildet, der als Ausfüllhilfe dient, aber nicht als Vorlage verwendet werden soll.

Beschreiben Sie nachfolgend die Verzerrungsaspekte jeder eingeschlossenen Studie (einschließlich der Beschreibung für jeden berücksichtigten Endpunkt). Erstellen Sie hierfür je Studie eine separate Version des nachfolgend dargestellten Bewertungsbogens.

Tabelle 4-29 (Anhang): Bewertungsbogen zur Beschreibung von Verzerrungsaspekten für Studie <Studienbezeichnung>

Studie: _____

Tabelle: Liste der für die Bewertung herangezogenen Quellen

Genaue Benennung der Quelle	Kürzel

A Verzerrungsaspekte auf Studienebene:

Einstufung als randomisierte Studie

☐ **ja** → Bewertung der Punkte 1 und 2 für randomisierte Studien

☐ **nein** → Bewertung der Punkte 1 und 2 für nicht randomisierte Studien

Angaben zum Kriterium:

_____

_____

1.

für randomisierte Studien: Adäquate Erzeugung der Randomisierungssequenz

☐ **ja** ☐ **unklar** ☐ **nein**

Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

_____

_____

für nicht randomisierte Studien: Zeitliche Parallelität der Gruppen

☐ **ja** ☐ **unklar** ☐ **nein**

Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

_____

_____

2.

für randomisierte Studien: Verdeckung der Gruppenzuteilung („allocation concealment“)

☐ **ja** ☐ **unklar** ☐ **nein**

Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

_____

_____

<u>für nicht randomisierte Studien:</u> Vergleichbarkeit der Gruppen bzw. adäquate Berücksichtigung von prognostisch relevanten Faktoren

☐ **ja** ☐ **unklar** ☐ **nein**

Angaben zum Kriterium; falls unklar oder nein, <u>obligate</u> Begründung für die Einstufung:

---

3. Verblindung von Patienten und behandelnden Personen

Patient:

☐ **ja** ☐ **unklar** ☐ **nein**

Angaben zum Kriterium; <u>obligate</u> Begründung für die Einstufung:

---

behandelnde bzw. weiterbehandelnde Personen:

☐ **ja** ☐ **unklar** ☐ **nein**

Angaben zum Kriterium; <u>obligate</u> Begründung für die Einstufung:

---

4. Ergebnisunabhängige Berichterstattung aller relevanten Endpunkte

☐ **ja** ☐ **unklar** ☐ **nein**

Angaben zum Kriterium; falls unklar oder nein, <u>obligate</u> Begründung für die Einstufung:

---

5. Keine sonstigen (endpunktübergreifenden) Aspekte, die zu Verzerrungen führen können

☐ **ja** ☐ **nein**

Angaben zum Kriterium; falls nein, <u>obligate</u> Begründung für die Einstufung:

Einstufung des Verzerrungspotenzials der Ergebnisse auf Studienebene (ausschließlich für randomisierte Studien durchzuführen):

☐ **niedrig** ☐ **hoch**

Begründung für die Einstufung:

B Verzerrungsaspekte auf Endpunktebene pro Endpunkt:

Endpunkt: _____

1. **Verblindung der Endpunkterheber**

   ☐ **ja** ☐ **unklar** ☐ **nein**

   Angaben zum Kriterium; obligate Begründung für die Einstufung:

2. **Adäquate Umsetzung des ITT-Prinzips**

   ☐ **ja** ☐ **unklar** ☐ **nein**

   Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

3. **Ergebnisunabhängige Berichterstattung dieses Endpunkts alleine**

   ☐ **ja** ☐ **unklar** ☐ **nein**

   Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

4. **Keine sonstigen (endpunktspezifischen) Aspekte, die zu Verzerrungen führen können**

   ☐ **ja** ☐ **nein**

   Angaben zum Kriterium; falls nein, obligate Begründung für die Einstufung:

Einstufung des Verzerrungspotenzials der Ergebnisse des Endpunkts (ausschließlich für randomisierte Studien durchzuführen):

☐ **niedrig** ☐ **hoch**

Begründung für die Einstufung:

Hinweis: Der nachfolgend dargestellte Bewertungsbogen mit Ausfüllhinweisen dient nur als Ausfüllhilfe für den Blankobogen. Er soll nicht als Vorlage verwendet werden.

Bewertungsbogen zur Beschreibung von Verzerrungsaspekten (Ausfüllhilfe)

Anhand der Bewertung der folgenden Kriterien soll das Ausmaß möglicher Ergebnisverzerrungen eingeschätzt werden (A: endpunktübergreifend; B: endpunktspezifisch).

A Verzerrungsaspekte auf Studienebene:

Einstufung als randomisierte Studie

☐ **ja** → Bewertung der Punkte 1 und 2 für randomisierte Studien

☐ **nein**: Aus den Angaben geht klar hervor, dass es keine randomisierte Zuteilung gab, oder die Studie ist zwar als randomisiert beschrieben, es liegen jedoch Anzeichen vor, die dem widersprechen (z. B. wenn eine alternierende Zuteilung erfolgte). Eine zusammenfassende Bewertung der Verzerrungsaspekte soll für nicht randomisierte Studien nicht vorgenommen werden.
→ Bewertung der Punkte 1 und 2 für nicht randomisierte Studien

Angaben zum Kriterium:

1.

für randomisierte Studien:

Adäquate Erzeugung der Randomisierungssequenz

☐ **ja**: Die Gruppenzuteilung erfolgte rein zufällig, und die Erzeugung der Zuteilungssequenz ist beschrieben und geeignet (z. B. computergenerierte Liste).

☐ **unklar**: Die Studie ist zwar als randomisiert beschrieben, die Angaben zur Erzeugung der Zuteilungssequenz fehlen jedoch oder sind ungenügend genau.

☐ **nein**: Die Erzeugung der Zuteilungssequenz war nicht adäquat.

Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

für nicht randomisierte Studien:

Zeitliche Parallelität der Gruppen

☐ **ja**: Die Gruppen wurden zeitlich parallel verfolgt.

☐ **unklar:** Es finden sich keine oder ungenügend genaue diesbezügliche Angaben.

☐ **nein**: Die Gruppen wurden nicht zeitlich parallel verfolgt.

Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

2.

für randomisierte Studien:

Verdeckung der Gruppenzuteilung („allocation concealment“)

☐ **ja**: Eines der folgenden Merkmale trifft zu:
- Zuteilung durch zentrale unabhängige Einheit (z. B. per Telefon oder Computer)
- Verwendung von für die Patienten und das medizinische Personal identisch aussehenden, nummerierten oder kodierten Arzneimitteln/Arzneimittelbehältern
- Verwendung eines seriennummerierten, versiegelten und undurchsichtigen Briefumschlags, der die Gruppenzuteilung beinhaltet

☐ **unklar**: Die Angaben der Methoden zur Verdeckung der Gruppenzuteilung fehlen oder sind ungenügend genau.

☐ **nein**: Die Gruppenzuteilung erfolgte nicht verdeckt.

Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

---

für nicht randomisierte Studien:

Vergleichbarkeit der Gruppen bzw. adäquate Berücksichtigung von prognostisch relevanten Faktoren

☐ **ja**: Eines der folgenden Merkmale trifft zu:
- Es erfolgte ein Matching bzgl. der wichtigen Einflussgrößen und es gibt keine Anzeichen dafür, dass die Ergebnisse durch weitere Einflussgrößen verzerrt sind.
- Die Gruppen sind entweder im Hinblick auf wichtige Einflussgrößen vergleichbar (siehe Baseline-Charakteristika), oder bestehende größere Unterschiede sind adäquat berücksichtigt worden (z. B. durch adjustierte Auswertung oder Sensitivitätsanalyse).

☐ **unklar**: Die Angaben zur Vergleichbarkeit der Gruppen bzw. zur Berücksichtigung von Einflussgrößen fehlen oder sind ungenügend genau.

☐ **nein**: Die Vergleichbarkeit ist nicht gegeben und diese Unterschiede werden in den Auswertungen nicht adäquat berücksichtigt.

Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

---

3. Verblindung von Patienten und behandelnden Personen

Patient:

☐ **ja:** Die Patienten waren verblindet.

☐ **unklar:** Es finden sich keine diesbezüglichen Angaben.

☐ **nein:** Aus den Angaben geht hervor, dass die Patienten nicht verblindet waren.

Angaben zum Kriterium; obligate Begründung für die Einstufung:

behandelnde bzw. weiterbehandelnde Personen:

☐ **ja:** Das behandelnde Personal war bzgl. der Behandlung verblindet. Wenn es, beispielsweise bei chirurgischen Eingriffen, offensichtlich nicht möglich ist, die primär behandelnde Person (z. B. Chirurg) zu verblinden, wird hier beurteilt, ob eine angemessene Verblindung der weiteren an der Behandlung beteiligten Personen (z. B. Pflegekräfte) stattgefunden hat.

☐ **unklar:** Es finden sich keine diesbezüglichen Angaben.

☐ **nein:** Aus den Angaben geht hervor, dass die behandelnden Personen nicht verblindet waren.

Angaben zum Kriterium; obligate Begründung für die Einstufung:

4. Ergebnisunabhängige Berichterstattung aller relevanten Endpunkte

Falls die Darstellung des Ergebnisses eines Endpunkts von seiner Ausprägung (d. h. vom Resultat) abhängt, können erhebliche Verzerrungen auftreten. Je nach Ergebnis kann die Darstellung unterlassen worden sein (a), mehr oder weniger detailliert (b) oder auch in einer von der Planung abweichenden Weise erfolgt sein (c).

Beispiele zu a und b:
- *Der in der Fallzahlplanung genannte primäre Endpunkt ist nicht / unzureichend im Ergebnisteil aufgeführt.*
- *Es werden (signifikante) Ergebnisse von vorab nicht definierten Endpunkten berichtet.*
- *Nur statistisch signifikante Ergebnisse werden mit Schätzern und Konfidenzintervallen dargestellt.*
- *Lediglich einzelne Items eines im Methodenteil genannten Scores werden berichtet.*

Beispiele zu c: Ergebnisgesteuerte Auswahl in der Auswertung verwendeter
- *Subgruppen*
- *Zeitpunkte/-räume*
- *Operationalisierungen von Zielkriterien (z. B. Wert zum Studienende anstelle der Veränderung zum Baseline-Wert; Kategorisierung anstelle Verwendung stetiger Werte)*
- *Distanzmaße (z. B. Odds Ratio anstelle der Risikodifferenz)*
- *Cut-off-points bei Dichotomisierung*
- *statistischer Verfahren*

Zur Einschätzung einer potenziell vorhandenen ergebnisgesteuerten Berichterstattung sollten folgende Punkte – sofern möglich – berücksichtigt werden:
- *Abgleich der Angaben der Quellen zur Studie (Studienprotokoll, Studienbericht, Registerbericht, Publikationen).*
- *Abgleich der Angaben im Methodenteil mit denen im Ergebnisteil. Insbesondere eine stark von der Fallzahlplanung abweichende tatsächliche Fallzahl ohne plausible und ergebnisunabhängige Begründung deutet auf eine selektive Beendigung der Studie hin.*
  Zulässige Gründe sind:
  - *erkennbar nicht ergebnisgesteuert, z. B. zu langsame Patientenrekrutierung*
  - *Fallzahladjustierung aufgrund einer verblindeten Zwischenauswertung anhand der Streuung der Stichprobe*
  - *geplante Interimanalysen, die zu einem vorzeitigen Studienabbruch geführt haben*
- *Prüfen, ob statistisch nicht signifikante Ergebnisse weniger ausführlich dargestellt sind.*
- *Ggf. prüfen, ob „übliche“ Endpunkte nicht berichtet sind.*

Anzumerken ist, dass Anzeichen für eine ergebnisgesteuerte Darstellung eines Endpunkts zu Verzerrungen der Ergebnisse der übrigen Endpunkte führen kann, da dort ggf. auch mit einer selektiven Darstellung gerechnet werden muss. Insbesondere bei Anzeichen dafür, dass die Ergebnisse einzelner Endpunkte selektiv nicht berichtet werden, sind Verzerrungen für die anderen Endpunkte möglich. Eine von der Planung abweichende selektive Darstellung des Ergebnisses eines Endpunkts führt jedoch nicht zwangsläufig zu einer Verzerrung der anderen Endpunkte; in diesem Fall ist die ergebnisgesteuerte Berichterstattung endpunktspezifisch unter Punkt B.3 (siehe unten) einzutragen.
Des Weiteren ist anzumerken, dass die Berichterstattung von unerwünschten Ereignissen üblicherweise ergebnisabhängig erfolgt (es werden nur Häufungen / Auffälligkeiten berichtet) und dies nicht zur Verzerrung anderer Endpunkte führt.

☐ **ja:** Eine ergebnisgesteuerte Berichterstattung ist unwahrscheinlich.

☐ **unklar:** Die verfügbaren Angaben lassen eine Einschätzung nicht zu.

☐ **nein:** Es liegen Anzeichen für eine ergebnisgesteuerte Berichterstattung vor, die das Verzerrungspotenzial aller relevanten Endpunkte beeinflusst.

Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

---

5. Keine sonstigen (endpunktübergreifenden) Aspekte, die zu Verzerrung führen können

z. B.
- *zwischen den Gruppen unterschiedliche Begleitbehandlungen außerhalb der zu evaluierenden Strategien*
- *intransparenter Patientenfluss*
- *Falls geplante Interimanalysen durchgeführt wurden, so sind folgende Punkte zu beachten:*
  - *Die Methodik muss exakt beschrieben sein (z. B. alpha spending approach nach O'Brien Fleming, maximale Stichprobengröße, geplante Anzahl und Zeitpunkte der Interimanalysen).*
  - *Die Resultate (p-Wert, Punkt- und Intervallschätzung) des Endpunktes, dessentwegen die Studie abgebrochen wurde, sollten adjustiert worden sein.*
  - *Eine Adjustierung sollte auch dann erfolgen, wenn die maximale Fallzahl erreicht wurde.*
  - *Sind weitere Endpunkte korreliert mit dem Endpunkt, dessentwegen die Studie abgebrochen wurde, so sollten diese ebenfalls adäquat adjustiert werden.*

☐ **ja**

☐ **nein**

Angaben zum Kriterium; falls nein, obligate Begründung für die Einstufung:

Einstufung des Verzerrungspotenzials der Ergebnisse auf Studienebene (ausschließlich für randomisierte Studien durchzuführen):

Die Einstufung des Verzerrungspotenzials der Ergebnisse erfolgt unter Berücksichtigung der einzelnen Bewertungen der vorangegangenen Punkte A.1 bis A.5. Eine relevante Verzerrung bedeutet hier, dass sich die Ergebnisse bei Behebung der verzerrenden Aspekte in ihrer Grundaussage verändern würden.

☐ **niedrig**: Es kann mit großer Wahrscheinlichkeit ausgeschlossen werden, dass die Ergebnisse durch diese endpunktübergreifenden Aspekte relevant verzerrt sind.

☐ **hoch**: Die Ergebnisse sind möglicherweise relevant verzerrt.

Begründung für die Einstufung:

B Verzerrungsaspekte auf Endpunktebene pro Endpunkt:

Die folgenden Punkte B.1 bis B.4 dienen der Einschätzung der endpunktspezifischen Aspekte für das Ausmaß möglicher Ergebnisverzerrungen. Diese Punkte sollten i. d. R. für jeden relevanten Endpunkt separat eingeschätzt werden (ggf. lassen sich mehrere Endpunkte gemeinsam bewerten, z. B. Endpunkte zu unerwünschten Ereignissen).

Endpunkt: _____

1. Verblindung der Endpunkterheber

Für den Endpunkt ist zu bestimmen, ob das Personal, welches die Zielkriterien erhoben hat, bzgl. der Behandlung verblindet war.
In manchen Fällen kann eine Verblindung auch gegenüber den Ergebnissen zu anderen Endpunkten (z. B. typischen unerwünschten Ereignissen) gefordert werden, wenn die Kenntnis dieser Ergebnisse Hinweise auf die verabreichte Therapie gibt und damit zu einer Entblindung führen kann.

☐ **ja:** Der Endpunkt wurde verblindet erhoben.

☐ **unklar:** Es finden sich keine diesbezüglichen Angaben.

☐ **nein:** Aus den Angaben geht hervor, dass keine verblindete Erhebung erfolgte.

Angaben zum Kriterium; obligate Begründung für die Einstufung:

---

2. Adäquate Umsetzung des ITT-Prinzips

Kommen in einer Studie Patienten vor, die die Studie entweder vorzeitig abgebrochen haben oder wegen Protokollverletzung ganz oder teilweise aus der Analyse ausgeschlossen wurden, so sind diese ausreichend genau zu beschreiben (Abbruchgründe, Häufigkeit und Patientencharakteristika pro Gruppe) oder in der statistischen Auswertung angemessen zu berücksichtigen (i. d. R. ITT-Analyse, siehe Äquivalenzstudien). Bei einer ITT(„intention to treat")-Analyse werden alle randomisierten Patienten entsprechend ihrer Gruppenzuteilung ausgewertet (ggf. müssen fehlende Werte für die Zielkriterien in geeigneter Weise ersetzt werden). Zu beachten ist, dass in Publikationen der Begriff ITT nicht immer in diesem strengen Sinne Verwendung findet. Es werden häufig nur die randomisierten Patienten ausgewertet, die die Therapie zumindest begonnen haben und für die mindestens ein Post-Baseline-Wert erhoben worden ist („full analysis set"). Dieses Vorgehen ist in begründeten Fällen Guideline-konform, eine mögliche Verzerrung sollte jedoch, insbesondere in nicht verblindeten Studien, überprüft werden. Bei Äquivalenz- und Nichtunterlegenheitsstudien ist es besonders wichtig, dass solche Patienten sehr genau beschrieben werden und die Methode zur Berücksichtigung dieser Patienten transparent dargestellt wird.

☐ **ja:** Eines der folgenden Merkmale trifft zu:

- Laut Studienunterlagen sind keine Protokollverletzer und Lost-to-follow-up-Patienten in relevanter Anzahl (z. B. Nichtberücksichtigungsanteil in der Auswertung < 5 %) aufgetreten, und es gibt keine Hinweise (z. B. diskrepante Patientenanzahlen in Flussdiagramm und Ergebnistabelle), die dies bezweifeln lassen.

- Die Protokollverletzer und Lost-to-follow-up-Patienten sind so genau beschrieben (Art, Häufigkeit und Charakteristika pro Gruppe), dass deren möglicher Einfluss auf die Ergebnisse abschätzbar ist (eigenständige Analyse möglich).
- Die Strategie zur Berücksichtigung von Protokollverletzern und Lost-to-follow-up-Patienten (u. a. Ersetzen von fehlenden Werten, Wahl der Zielkriterien, statistische Verfahren) ist sinnvoll angelegt worden (verzerrt die Effekte nicht zugunsten der zu evaluierenden Behandlung).

☐ **unklar**: Aufgrund unzureichender Darstellung ist der adäquate Umgang mit Protokollverletzern und Lost-to-follow-up-Patienten nicht einschätzbar.

☐ **nein**: Keines der unter „ja“ genannten drei Merkmale trifft zu.

Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

---

3. Ergebnisunabhängige Berichterstattung dieses Endpunkts alleine

Beachte die Hinweise zu Punkt A.4!

☐ **ja:** Eine ergebnisgesteuerte Berichterstattung ist unwahrscheinlich.

☐ **unklar:** Die verfügbaren Angaben lassen eine Einschätzung nicht zu.

☐ **nein:** Es liegen Anzeichen für eine ergebnisgesteuerte Berichterstattung vor.

Angaben zum Kriterium; falls unklar oder nein, obligate Begründung für die Einstufung:

---

4. Keine sonstigen (endpunktspezifischen) Aspekte, die zu Verzerrungen führen können

z. B.
- *relevante Dateninkonsistenzen innerhalb der oder zwischen Studienunterlagen*
- *unplausible Angaben*
- *Anwendung inadäquater statistischer Verfahren*

☐ **ja**

☐ **nein**

Angaben zum Kriterium; falls nein, obligate Begründung für die Einstufung:

Einstufung des Verzerrungspotenzials der Ergebnisse des Endpunkts (ausschließlich für randomisierte Studien durchzuführen):

Die Einstufung des Verzerrungspotenzials erfolgt unter Berücksichtigung der einzelnen Bewertungen der vorangegangenen endpunktspezifischen Punkte B.1 bis B.4 sowie der Einstufung des Verzerrungspotenzials auf Studienebene. Falls die endpunktübergreifende Einstufung mit „hoch“ erfolgte, ist das Verzerrungspotenzial für den Endpunkt i. d. R. auch mit „hoch“ einzuschätzen. Eine relevante Verzerrung bedeutet hier, dass sich die Ergebnisse bei Behebung der verzerrenden Aspekte in ihrer Grundaussage verändern würden.

☐ **niedrig**: Es kann mit großer Wahrscheinlichkeit ausgeschlossen werden, dass die Ergebnisse für diesen Endpunkt durch die endpunktspezifischen sowie endpunktübergreifenden Aspekte relevant verzerrt sind.

☐ **hoch**: Die Ergebnisse sind möglicherweise relevant verzerrt.

Begründung für die Einstufung:

Tragende Gründe

Gemeinsamer Bundesausschuss

zum Beschluss des Gemeinsamen Bundesausschusses über die Einleitung eines Stellungnahmeverfahrens zur Änderung der Verfahrensordnung: Änderung der Modulvorlage in der Anlage II zum 5. Kapitel

Vom 17. Juni 2021

Inhalt

1. Rechtsgrundlage

Der Gemeinsame Bundesausschuss (G-BA) hat gemäß § 91 Absatz 4 Satz 1 Nummer 1 SGB V eine Verfahrensordnung zu beschließen, in der er insbesondere methodische Anforderungen an die wissenschaftliche sektorenübergreifende Bewertung des Nutzens, der Notwendigkeit und der Wirtschaftlichkeit von Maßnahmen als Grundlage für Beschlüsse sowie die Anforderungen an den Nachweis der fachlichen Unabhängigkeit von Sachverständigen und anzuhörenden Stellen, die Art und Weise der Anhörung und deren Auswertung regelt. Die Verfahrensordnung bedarf gemäß § 91 Absatz 4 Satz 2 SGB V der Genehmigung des Bundesministeriums für Gesundheit. Mit Beschluss vom 20. Januar 2011 hat der G-BA ein 5. Kapitel in die Verfahrensordnung eingefügt, in dem das Nähere zum Verfahren über die Bewertung des Zusatznutzens von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V geregelt ist.

2. Eckpunkte der Entscheidung

Der G-BA hat in seiner Sitzung am 17. Juni 2021 beschlossen, ein Stellungnahmeverfahren zur Änderung der Modulvorlagen in der Anlage II zum 5. Kapitel der Verfahrensordnung (VerfO) einzuleiten.

Gemäß 1. Kapitel § 8 Absatz 2 Satz 1 Buchstabe b VerfO kann das Plenum im Einzelfall beschließen, dass zu Entscheidungen, bei denen kein gesetzlich eingeräumtes Stellungnahmerecht besteht, ebenfalls Stellungnahmen einzuholen sind.

Mit dem vorliegenden Beschlussentwurf sollen Anpassungen der Anlage II.6 (Modul 4 – Medizinischer Nutzen und medizinischer Zusatznutzen, Patientengruppen mit therapeutisch bedeutsamem Zusatznutzen) zum 5. Kapitel der VerfO vorgenommen werden, welche durch Änderungen der methodischen Anforderungen an die Dossiererstellung in Verbindung mit dem bisherigen Vorgehen und den Erfahrungen des G-BA mit der Nutzenbewertung nach § 35a SGB V erforderlich geworden sind. Zur Abbildung des dazu aktuell bestehenden wissenschaftlichen Diskurses bedient sich der G-BA im Rahmen seines Entscheidungsermessens ausnahmsweise ergänzend des Stellungnahmeverfahrens.

Die Änderungen betreffen in dem Abschnitt 4.3.1 (Ergebnisse randomisierter kontrollierter Studien mit dem zu bewertenden Arzneimittel) den Unterabschnitt 4.3.1.3.1 (<Endpunkt xxx> – RCT) der Anlagen II.6 zum 5. Kapitel der VerfO. Diesbezüglich sollen Konkretisierungen zur Ergebnisdarstellung von patientenberichteten Endpunkten vorgenommen werden, wie das Vorgehen zur Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen erfolgen soll.

Gemäß dem aktuellen methodischen Vorgehen des IQWiG (Methodenpapier 6.0, veröffentlicht am 05.11.2020) erachtet das IQWiG für patientenberichtete Endpunkte eine Responseschwelle für Responderanalysen von mindestens 15 % der Skalenspannweite eines Instrumentes (bei post hoc durchgeführten Analysen von genau 15 % der Skalenspannweite) als notwendig, um eine für Patienten spürbare Veränderung hinreichend sicher abzubilden.

Hintergrund ist, dass sich bei Responderanalysen auf Basis eines Responsekriteriums im Sinne einer individuellen Minimal important Difference (MID) vermehrt methodische Probleme offenbart haben.

Demnach zeigen systematische Zusammenstellungen empirisch ermittelter MIDs, dass zu einzelnen Instrumenten häufig eine Vielzahl von MIDs publiziert werden, die innerhalb eines Erhebungsinstruments große Spannweiten haben können[1, 2, 3, 4]. Ursachlich hierfür können unter anderem die in den Studien eingesetzten unterschiedliche Anker, Beobachtungsperioden oder analytische Methoden sein[5, 4, 6]. Gleichzeitig ist eine anhand methodischer Qualitätskriterien begründete Auswahl empirisch ermittelter MIDs für die Nutzenbewertung derzeit nicht zu treffen[7, 5, 8].

Neben den methodischen Faktoren beruht ein anderer Teil der Variabilität von MIDs auf ihrer Abhängigkeit von Charakteristika der Patientenpopulation, in der das Instrument eingesetzt wird, sowie weiteren Kontextfaktoren. So können der Schweregrad der Erkrankung, die Art der eingesetzten Intervention oder die Frage, ob die Patientinnen und Patienten eine Verbesserung oder Verschlechterung ihrer Erkrankung erfahren, Einfluss auf die MID haben[9]. Der Umgang mit diesem Teil der Variabilität von MIDs ist ungeklärt.

Insgesamt gehen die genannten Limitationen bei Responderanalysen auf Basis eines Responsekriteriums im Sinne einer MID mit wesentlichen Unsicherheiten bei der Interpretation der klinischen Relevanz des beobachteten Effektes einher.

Laut IQWiG (Methodenpapier 6.0, veröffentlicht am 05.11.2020) wurde ein Wert von 15 % der Spannweite der jeweiligen Skalen als plausibler Schwellenwert für eine hinreichend sicher spürbare Veränderung identifiziert.

---

[1] Carrasco-Labra A, Devji T, Qasim A, Phillips M, Devasenapathy N, Zeraatkar D et al. Interpretation of patient-reported outcome measures: an inventory of over 3000 minimally important difference estimates and an assessment of their credibility. Cochrane Database Syst Rev 2018; (9 Suppl 1): 135-136.

[2] Çelik D, Çoban Ö, Kılıçoğlu Ö. Minimal clinically important difference of commonly used hip-, knee-, foot-, and ankle-specific questionnaires: a systematic review. J Clin Epidemiol 2019; 113: 44-57.

[3] Hao Q, Devji T, Zeraatkar D, Wang Y, Qasim A, Siemieniuk RAC et al. Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation. BMJ Open 2019; 9(2): e028777.

[4] Nordin A, Taft C, Lundgren-Nilsson A, Dencker A. Minimal important differences for fatigue patient reported outcome measures: a systematic review. BMC Med Res Methodol 2016; 16: 62.

[5] Devji T, Guyatt GH, Lytvyn L, Brignardello-Petersen R, Foroutan F, Sadeghirad B et al. Application of minimal important differences in degenerative knee disease outcomes: a systematic review and case study to inform BMJ Rapid Recommendations. BMJ Open 2017; 7(5): e015587.

[6] Ousmen A, Touraine C, Deliu N, Cottone F, Bonnetain F, Efficace F et al. Distribution- and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review. Health Qual Life Outcomes 2018; 16(1): 228

[7] Devji T, Carrasco-Labra A, Lytvyn L, Johnston B, Ebrahim S, Furukawa T et al. A new tool to measure credibility of studies determining minimally important difference estimates. Cochrane Database Syst Rev 2017; (9 Suppl 1): 58.

[8] Johnston BC, Ebrahim S, Carrasco-Labra A, Furukawa TA, Patrick DL, Crawford MW et al. Minimally important difference estimates and methods: a protocol. BMJ Open 2015; 5(10): e007953.

[9] Alma H, De Jong C, Jelusic D, Wittmann M, Schuler M, Kollen B et al. Baseline health status and setting impacted minimal clinically important differences in COPD: an exploratory study. J Clin Epidemiol 2019; 116: 49-61.

Die mit dem vorliegenden Beschlussentwurf vorgesehene Anpassung der Anlage II.6 zum 5. Kapitel der VerfO soll daher sicherstellen, dass in Responderanalysen im Rahmen der Nutzenbewertung geeignete Responseschwellen eingesetzt werden, die eine für die Patientinnen und Patienten spürbare Veränderungen abbilden und die Gefahr einer ergebnisgesteuerten Berichterstattung minimieren, womit diesbezügliche Unsicherheiten bei der Interpretation der klinischen Relevanz des beobachteten Effektes verhindert werden sollen. Vor dem Hintergrund der Änderungen der methodischen Anforderungen an die Dossiererstellung in Verbindung mit dem bisherigen Vorgehen und den Erfahrungen des G-BA mit der Nutzenbewertung nach § 35a SGB V sollen zudem Unsicherheiten der pharmazeutischen Unternehmer in der Dossiererstellung im Hinblick auf die Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen vermieden werden. Gleichzeitig sollen mit den Änderungen in der Modulvorlage Lücken in der Ergebnisdarstellung zu patientenberichteten Endpunkten, welche teilweise erst im Rahmen des Stellungnahmeverfahrens aufgeklärt werden können, vermieden werden.

Mit dem vorliegenden Beschlussentwurf zur Anpassungen der Anlage II.6 zum 5. Kapitel der VerfO wird der Unterabschnitt 4.3.1.3.1 (<Endpunkt xxx> – RCT) daher wie folgt geändert:

Es wird ergänzt, wie das Vorgehen zur Auswertung von Responderanalysen mittels klinischer Relevanzschwellen bei komplexen Skalen im Rahmen der Dossiererstellung erfolgen soll:

1. Falls in einer Studie Responderanalysen unter Verwendung einer MID präspezifiziert sind und das Responsekriterium mindestens 15 % der Skalenspannweite des verwendeten Erhebungsinstruments entspricht, sind diese Responderanalysen des Responsekriteriums für die Bewertung darzustellen.
2. Falls präspezifiziert Responsekriterien im Sinne einer MID unterhalb von 15 % der Skalenspannweite liegen, bestehen in diesen Fällen und solchen, in denen gar keine Responsekriterien präspezifiziert wurden, aber stattdessen Analysen kontinuierlicher Daten zur Verfügung stehen, verschiedene Möglichkeiten. Entweder können die Analysen der kontinuierlichen Daten dargestellt werden, für die Relevanzbewertung ist dabei auf ein allgemeines statistisches Maß in Form von standardisierten Mittelwertdifferenzen (SMDs, in Form von Hedges' g) zurückzugreifen. Dabei ist eine Irrelevanzschwelle von 0,2 zu verwenden: Liegt das zum Effektschätzer korrespondierende Konfidenzintervall vollständig oberhalb dieser Irrelevanzschwelle, wird davon ausgegangen, dass die Effektstärke nicht in einem sicher irrelevanten Bereich liegt. Dies soll gewährleisten, dass der Effekt hinreichend sicher mindestens als klein angesehen werden kann. Alternativ können post hoc spezifizierte Analysen mit einem Responsekriterium von genau 15 % der Skalenspannweite dargestellt werden.

3. Bürokratiekostenermittlung

Durch den vorgesehenen Beschluss entstehen keine neuen bzw. geänderten Informationspflichten für Leistungserbringer im Sinne von Anlage II zum 1. Kapitel VerfO und dementsprechend keine Bürokratiekosten.

4. Verfahrensablauf

Der Unterausschuss Arzneimittel hat zur Vorbereitung einer Überarbeitung der VerfO zur Änderung der Anlage II.6 zum 5. Kapitel und Erstellung einer Beschlussempfehlung zur Einleitung eines Stellungnahmeverfahrens zur Änderung der VerfO im 5. Kapitel die Arbeitsgruppe Entscheidungsgrundlagen beauftragt.

Der Unterausschuss Arzneimittel hat in seiner Sitzung am 8. Juni 2021 über die Änderungen im 5. Kapitel der VerfO beraten und die Beschlussvorlage über die Einleitung eines Stellungnahmeverfahrens zur Änderung der VerfO im 5. Kapitel konsentiert.

Die Beschlussvorlage wurde der Arbeitsgruppe Geschäftsordnung-Verfahrensordnung übersandt, die am 10. Juni 2021 schriftlich über die Beschlussunterlagen abgestimmt und diese an das Plenum des Gemeinsamen Bundesausschusses zur Beschlussfassung nach 1. Kapitel § 8 Absatz 2 Satz 1 Buchstabe b VerfO weitergeleitet hat.

Das Plenum des Gemeinsamen Bundesausschusses hat am 17. Juni 2021 über die Beschlussempfehlungen zur Einleitung eines Stellungnahmeverfahrens zur Änderung der VerfO im 5. Kapitel beraten und die Einleitung eines Stellungnahmeverfahrens zur Änderung der VerfO im 5. Kapitel – Änderungen der Modulvorlagen in der Anlage II beschlossen.

Zeitlicher Beratungsverlauf

Sitzung	**Datum**	**Beratungsgegenstand**
AG Entscheidungsgrundlagen	3. Mai 2021 31. Mai 2021	Beratung über die Änderungen der Anlage II.6 zum 5. Kapitel der Verfahrensordnung und Erstellung einer Beschlussempfehlung zur Einleitung eines diesbezüglichen Stellungnahmeverfahrens
Unterausschuss Arzneimittel	8. Juni 2021	Beratung und Konsentierung der Beschlussvorlage zur Einleitung des Stellungnahmeverfahrens hinsichtlich der Änderung der Anlage II.6 zum 5. Kapitel der Verfahrensordnung
AG Geschäftsordnung-Verfahrensordnung	10. Juni 2021	Schriftliche Abstimmung über die Beschlussvorlage
Plenum	17. Juni 2021	Beschlussfassung zur Einleitung des Stellungnahmeverfahrens hinsichtlich der Änderung der Anlage II.6 zum 5. Kapitel der Verfahrensordnung

Zum Zeitpunkt der Einleitung des Stellungnahmeverfahrens stellen die vorliegenden Tragenden Gründe den aktuellen Stand der Zusammenfassenden Dokumentation dar, welche

den stellungnahmeberechtigten Organisationen zur Verfügung zu stellen sind (1. Kapitel § 10 Absatz 2 VerfO).

Als Frist zur Stellungnahme ist ein Zeitraum von 4 Wochen vorgesehen.

Eine Stellungnahme ist durch Literatur (z. B. relevante Studien) zu begründen. Die zitierte Literatur ist obligat im Volltext inklusive einem standardisierten und vollständigen Literatur- bzw. Anlagenverzeichnis der Stellungnahme beizufügen. Nur Literatur, die im Volltext beigefügt ist, kann berücksichtigt werden.

Mit Abgabe einer Stellungnahme erklärt sich der Stellungnehmer einverstanden, dass diese in den Tragenden Gründen bzw. in der Zusammenfassenden Dokumentation wiedergegeben werden kann. Diese Dokumente werden jeweils mit Abschluss der Beratungen im Gemeinsamen Bundesausschuss erstellt und in der Regel der Öffentlichkeit via Internet zugänglich gemacht.

Stellungnahmeberechtigte

Bezüglich Änderungen der Arzneimittel-Richtlinie aufgrund von Beschlüssen zur frühen Nutzenbewertung von Arzneimitteln sind die Sachverständigen der medizinischen und pharmazeutischen Wissenschaft und Praxis sowie die für die Wahrnehmung der wirtschaftlichen Interessen gebildeten maßgeblichen Spitzenorganisationen der pharmazeutischen Unternehmer, die betroffenen pharmazeutischen Unternehmer, die Berufsvertretungen der Apotheker und die maßgeblichen Dachverbände der Ärztegesellschaften der besonderen Therapierichtungen auf Bundesebene gemäß § 35a Absatz 3 Satz 2 i.V.m. § 92 Absatz 3a SGB V stellungnahmeberechtigt. Unter entsprechender Anwendung dieser Stellungnahmerechte wird der Beschlussentwurf zur Änderung der Anlage II.6 zum 5. Kapitel der VerfO den folgenden Organisationen sowie den Verbänden der pharmazeutischen Unternehmen mit der Bitte um Weiterleitung zugesendet.

Folgende Organisationen werden angeschrieben:

Organisation	**Straße**	**Ort**
Bundesverband der Pharmazeutischen Industrie e. V. (BPI)	Friedrichstr. 148	10117 Berlin
Verband Forschender Arzneimittelhersteller e. V. (vfa)	Hausvogteiplatz 13	10117 Berlin
Bundesverband der Arzneimittel-Importeure e. V. (BAI)	EurimPark 8	83416 Saaldorf-Surheim
Bundesverband der Arzneimittel-Hersteller e. V. (BAH)	Friedrichstr. 134	10117 Berlin
Biotechnologie-Industrie-Organisation Deutschland e. V. (BIO Deutschland e. V.)	Schützenstraße 6a	10117 Berlin
Pro Generika e. V.	Unter den Linden 32 - 34	10117 Berlin

Arzneimittelkommission der Deutschen Ärzteschaft (AkdÄ)	Herbert-Lewin-Platz 1	10623 Berlin
Arzneimittelkommission der Deutschen Zahnärzteschaft (AK-Z) c/o Bundeszahnärztekammer	Chausseestr. 13	10115 Berlin
Bundesvereinigung Deutscher Apothekerverbände e. V. (ABDA)	Heidestr. 7	10557 Berlin
Deutscher Zentralverein Homöopathischer Ärzte e. V.	Axel-Springer-Str. 54b	10117 Berlin
Gesellschaft Anthroposophischer Ärzte e. V.	Herzog-Heinrich-Str. 18	80336 München
Gesellschaft für Phytotherapie e. V.	Postfach 10 08 88	18055 Rostock

Darüber hinaus wird die Einleitung des Stellungnahmeverfahrens im Bundesanzeiger bekanntgemacht.

Betroffene pharmazeutische Unternehmen und Organisationen, die nicht Mitglieder der oben genannten Verbände sind, erhalten den Entwurf, die Technische Anlage sowie die Tragenden Gründe bei der Geschäftsstelle des Gemeinsamen Bundesausschusses.

Der Beschluss und die Tragenden Gründe können auf den Internetseiten des Gemeinsamen Bundesausschusses unter www.g-ba.de eingesehen werden.

Die mündliche Anhörung wird am 28. September 2021 um 10:00 Uhr in der Geschäftsstelle des G-BA durchgeführt. Zeitgleich mit der Einreichung der schriftlichen Stellungnahme ist sich zu der mündlichen Anhörung anzumelden, sofern an dieser teilgenommen werden möchte.

Berlin, den 17. Juni 2021

Gemeinsamer Bundesausschuss
gemäß § 91 SGB V
Der Vorsitzende

Prof. Hecken