

# Mündliche Anhörung



Gemeinsamer  
Bundesausschuss

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 Verfahrensordnung 5. Kapitel, Änderung  
der Modulvorlage in der Anlage II**

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



**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 probenhalber 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 Parameter-raum 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  $\geq$  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 Nutzensaussage, 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 Atemwegliga, 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 Atemwegliga): 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 Atemwegliga):** 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

Petitem, 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 verminten 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