

Schwellenwerte für Responderanalysen bei komplexen Skalen in der Nutzenbewertung – der 15 %-Ansatz

A decorative horizontal bar spanning the width of the page, composed of 18 small squares in various shades of blue and grey. The word 'ARBEITSPAPIER' is centered in a dark blue rectangle that covers the middle squares of this bar.

ARBEITSPAPIER

Projekt: GA25-04

Version: 1.0

Stand: 13.07.2026

IQWiG-Berichte – Nr. 2286

DOI: 10.60584/GA25-04

Impressum

Herausgeber

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Thema

Schwellenwerte für Responderanalysen bei komplexen Skalen in der Nutzenbewertung – der 15 %-Ansatz

Auftraggeber

Bearbeitung im Rahmen des Generalauftrags

Interne Projektnummer

GA25-04

DOI-URL

<https://doi.org/10.60584/GA25-04>

Anschrift des Herausgebers

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Siegburger Str. 237
50679 Köln

Tel.: +49 221 35685-0
Fax: +49 221 35685-1
E-Mail: info@iqwig.de
Internet: www.iqwig.de

ISSN: 1864-2500

Zitiervorschlag

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Schwellenwerte für Responderanalysen bei komplexen Skalen in der Nutzenbewertung – der 15 %-Ansatz ; Arbeitspapier [online]. 2026 [Zugriff: TT.MM.JJJJ]. URL: <https://doi.org/10.60584/GA25-04>.

Schlagwörter

Patientenberichtete Endpunkte

Keywords

Patient Reported Outcomes

Dieses Arbeitspapier wurde ohne die Beteiligung externer Sachverständiger erstellt.

Für die Inhalte des Berichts ist allein das IQWiG verantwortlich.

Mitarbeiterinnen und Mitarbeiter des IQWiG

- Ulrike Seay
- Katrin Nink
- Stefanie Thomas
- Beate Wieseler
- André Wilmer

Inhaltsverzeichnis

	Seite
Tabellenverzeichnis	v
Abbildungsverzeichnis	vi
Abkürzungsverzeichnis.....	vii
Kurzfassung	viii
1 Hintergrund.....	1
2 Fragestellung.....	2
3 Projektverlauf.....	3
4 Methoden	4
5 Ergebnisse	6
5.1 Ergebnis der Informationsbeschaffung.....	6
5.2 Ergebnisse zu MIDs aus den in die systematischen Übersichtsarbeiten eingeschlossenen Studien.....	7
6 Diskussion	12
6.1 Anhörung zum Verfahrensvorschlag	12
6.2 Internationale Rezeption des 15 %-Ansatzes	15
7 Fazit	16
8 Literatur	17
Anhang A Bibliografische Recherche.....	21
Anhang B Tabellen zu den identifizierten MIDs in den einzelnen systematischen Übersichtsarbeiten	22
Anhang C Spezifische Instrumente	27
C.1 EORTC QLQ-C30 und seine Zusatzmodule: Eignung der Responseschwellen von 10 Punkten für Auswertungen der EORTC-Module	27
C.2 Herleitung der Responsekriterien für den Physical Component Summary (PCS) und Mental Component Summary (MCS) des SF-36.....	30
C.3 PROMIS / Item-Banken.....	33

Tabellenverzeichnis

	Seite
Tabelle 1: Spannweite der empirisch erhobenen MID's im Verhältnis zur Spannweite der jeweils verwendeten Skala in %.....	9
Tabelle 2: Zusammenfassung der MID's für den WOMAC aus 8 in die Übersichtsarbeit Doganay eingeschlossenen Studien	10
Tabelle 3: Zusammenfassung der MID's der aus der Übersichtsarbeit Doganay berücksichtigten Studien	22
Tabelle 4: Zusammenfassung der MID's der aus der Übersichtsarbeit St. Pierre berücksichtigten Studien	23
Tabelle 5: Zusammenfassung der MID's der aus der Übersichtsarbeit Hao berücksichtigten Studien	23
Tabelle 6: Zusammenfassung der MID's der aus der Übersichtsarbeit Ebrahim berücksichtigten Studien	24
Tabelle 7: Zusammenfassung der MID's der aus der Übersichtsarbeit Nordin berücksichtigten Studien	24
Tabelle 8: Zusammenfassung der MID's der aus der Übersichtsarbeit Jayadevappa berücksichtigten Studien	25
Tabelle 9: Zusammenfassung der MID's der aus der Übersichtsarbeit Alma berücksichtigten Studien	26
Tabelle 10: Zusammenfassung der MID's der aus der Übersichtsarbeit Ousmen berücksichtigten Studien	26
Tabelle 11: Bedeutung der unterschiedlichen Responseschwellen 10 Punkte / 15 Punkte für die Skalen des EORTC QLQ-C30	28
Tabelle 12: Anzahl Items pro Domäne und Gewichtung der Domänen in den Component Summaries PCS und MCS.....	31
Tabelle 13: Minima und Maxima des PCS und MCS des SF-36 in verschiedenen Szenarien...	32

Abbildungsverzeichnis

	Seite
Abbildung 1: Ergebnis der fokussierten Informationsbeschaffung nach systematischen Übersichtsarbeiten zu MIDs	6

Abkürzungsverzeichnis

Abkürzung	Bedeutung
CAT	computer-adaptives Testen
COPD	Chronic obstructive pulmonary Disease (chronisch obstruktive Lungenerkrankung)
EORTC	European Organisation for Research and Treatment of Cancer
G-BA	Gemeinsamer Bundesausschuss
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
MCS	Mental Component Summary
MID	Minimal important Difference
PCS	Physical Component Summary
QLQ-C30	Quality of Life Questionnaire – Core 30
PRO	Patient Reported Outcome (patientenberichteter Endpunkt)
RCT	Randomized controlled Trial (randomisierte kontrollierte Studie)
ROC	Receiver Operating Characteristic
SF-36	Short Form 36
SGRQ	St. George's Respiratory Questionnaire
WOMAC	Western Ontario and McMaster Osteoarthritis Index

Kurzfassung

Im Jahr 2020 hat das IQWiG ein neues Verfahren für die Festlegung von Schwellenwerten für Responderanalysen bei komplexen Skalen, insbesondere solchen zur Bewertung patientenberichteter Endpunkte (PROs), vorgeschlagen und zur Anhörung gestellt. Seit Veröffentlichung der Allgemeinen Methoden Version 6.0 wird dieses Verfahren, der sogenannte 15 %-Ansatz, angewendet.

Hintergrund für die Beschäftigung mit Responseschwellen war, dass in den vorangegangenen Jahren für die Bewertung medizinischer Interventionen vermehrt Responderanalysen durchgeführt wurden, die ein Responsekriterium im Sinne einer individuellen Minimal important Difference (MID) heranziehen. Die Bestimmung einer MID ist allerdings mit methodischen Problemen verbunden und für einzelne Instrumente wird häufig eine Vielzahl von MIDs publiziert. Dies bedeutet, dass sich für Skalen keine einheitlichen MIDs identifizieren lassen, sondern diese von der jeweils betrachteten Patientengruppe, ihrer Erkrankung, dem Schweregrad dieser Erkrankung aber auch der Methodik der MID-Bestimmung abhängen. Auch wenn es mittlerweile Ansätze gibt, die Qualität von Validierungsstudien zu MIDs zu bewerten, zeigt sich, dass in Publikationen von Studien zur Ermittlung einer MID häufig wesentliche Anteile der Methodik bzw. der Kriterien für eine Bewertung nicht berichtet werden.

Das Ziel war daher, eine generische Responseschwelle zu identifizieren, die hinreichend sicher eine für Patientinnen und Patienten spürbare Veränderung darstellt.

Zu diesem Zweck wurden mittels einer fokussierten Suche systematische Übersichtsarbeiten zu MIDs identifiziert und die dort berichteten MIDs zusammengestellt. Insgesamt wurden aus 8 von 16 identifizierten systematische Übersichtsarbeiten die dort berichteten MIDs extrahiert. Diese umfassen einen breiten Themenbereich wie orthopädische und onkologische Indikationen, Pädiatrie oder Instrumente zur Erhebung gesundheitsbezogener Lebensqualität oder aber spezifische Konzepte (Fatigue) und Erkrankungen (chronisch obstruktiver Lungenerkrankung, Rheuma). Die identifizierten MIDs wurden in einem weiteren Schritt in Bezug zur Skalenspannweite des jeweiligen Erhebungsinstruments gesetzt.

Als Ergebnis wurde ein Wert von 15 % der Skalenspannweite der jeweiligen Skala als Responsekriterium festgelegt. Dieses Responsekriterium wird als plausibler Wert für eine relativ kleine aber hinreichend sicher im spürbaren Bereich liegende individuelle Änderung auf einer Skala angesehen. Hierbei handelt es sich explizit nicht um ein Responsekriterium im Sinne einer individuellen MID. Auch wenn die 15 %-Schwelle eine relativ kleine Veränderung abbilden soll, kann sie in vielen Fällen oberhalb identifizierter MIDs liegen.

Das IQWiG hat das Verfahren im Jahr 2020 erstmals mit seinen Allgemeinen Methoden Version 6.0 vorgestellt und im Rahmen des Stellungnahmeverfahrens sowie der mündlichen Erörterung mit der wissenschaftlichen Fachwelt diskutiert. Im Jahr 2021 hat der Gemeinsame Bundesausschuss (G-BA) die neue Schwelle in die Dossiervorlagen für die Bewertung des Zusatznutzens von neuen Arzneimitteln übernommen. Auch der G-BA hat hierzu ein Stellungnahmeverfahren und eine mündliche Anhörung durchgeführt. Seitdem hat sich die Anwendung der 15 %-Schwelle in dem nationalen Verfahren der Nutzenbewertung neuer Arzneimittel etabliert. Das vorliegende Arbeitspapier hat zum Ziel, die Methodik zur Herleitung der generischen Responseschwelle, die Diskussion im Rahmen des Stellungnahmeverfahrens zu den Allgemeinen Methoden Version 6.0 sowie weitere Festlegungen für spezifische Instrumente in einem Dokument zusammenzubringen. Mit der Zusammenfassung sollen die vorliegenden Informationen einem breiteren Leserkreis zugänglich gemacht werden.

1 Hintergrund

Im Jahr 2020 hat das IQWiG ein neues Verfahren für die Festlegung von Schwellenwerten für Responderanalysen bei komplexen Skalen, insbesondere solchen zur Bewertung patientenberichteter Endpunkte (PROs), vorgeschlagen und zur Anhörung gestellt. Seit Veröffentlichung der Allgemeinen Methoden Version 6.0 [1] wird dieses Verfahren, der sogenannte 15 %-Ansatz, angewendet.

Hintergrund für die Beschäftigung mit Responseschwellen war, dass in den vorangegangenen Jahren für die Bewertung medizinischer Interventionen vermehrt Responderanalysen durchgeführt wurden, die ein Responsekriterium im Sinne einer individuellen Minimal important Difference (MID) heranziehen. Unter einer MID wird dabei der kleinste Unterschied des Scores eines Instruments verstanden, der für einen Patienten oder einer Patientin spürbar ist und als Verbesserung oder Verschlechterung bewertet wird [2,3]. Die Bestimmung einer MID ist allerdings mit methodischen Problemen verbunden und für einzelne Instrumente wird häufig eine Vielzahl von MIDs publiziert [4]. Dies bedeutet, dass sich für Skalen keine einheitlichen MIDs identifizieren lassen, sondern diese von der jeweils betrachteten Patientengruppe, ihrer Erkrankung, dem Schweregrad dieser Erkrankung aber auch der Methodik der MID-Bestimmung abhängen. Auch wenn es mittlerweile Ansätze gibt, die Qualität von Validierungsstudie zu MIDs zu bewerten, zeigt sich, dass in Publikationen von Studien zur Ermittlung einer MID häufig wesentliche Informationen bzw. Kriterien zur Bewertung der Methodik nicht berichtet werden [5].

Ziel der damaligen Untersuchung war daher, eine generische Responseschwelle zu identifizieren, die zwar einerseits eher klein ist, sich aber gleichzeitig hinreichend sicher im für die Mehrzahl der Patientinnen und Patienten spürbaren Bereich befindet. Dabei wurde als Startpunkt eine generische Schwelle von 10 % der Skalenspannweite des Instruments diskutiert, da diese in der wissenschaftlichen Diskussion verschiedentlich als eine „grobe“ und kontextunabhängige Annäherung an eine MID diskutiert wurde [6,7]. Um den neuen Ansatz zur Festlegung eines generischen Responsekriteriums zu entwickeln, hat das IQWiG mittels einer fokussierten Suche systematische Übersichtsarbeiten zu MIDs identifiziert, die dort berichteten MIDs zusammengestellt und ins Verhältnis zur Skalenspannweite des jeweiligen Instruments gesetzt.

2 Fragestellung

Ziel der damaligen Untersuchung war es, eine generische Responseschwelle für Responderanalysen bei komplexen Skalen, insbesondere solchen zur Erhebung patientenberichteter Endpunkte, zu identifizieren, die hinreichend sicher eine für Patientinnen und Patienten spürbare Veränderung darstellt. Diese Schwelle soll zwar eine eher kleine Änderung darstellen, aber explizit keine „generische“ MID identifizieren. Daher kann sie oberhalb bereits identifizierter MIDs eines Instruments liegen.

3 Projektverlauf

Mit den Allgemeinen Methoden Version 6.0 hat das IQWiG im Jahr 2020 ein neues Verfahren für die Festlegung von Schwellenwerten für Responderanalysen bei komplexen Skalen, insbesondere solchen zur Erhebung patientenberichteter Endpunkte, veröffentlicht [1] und im Rahmen des Stellungnahmeverfahrens sowie der mündlichen Anhörung mit der Fachwelt diskutiert [8]. Seitdem wird die dort festgelegte Responseschwelle von 15 % der Skalenspannweite des jeweiligen Instruments angewendet. Der Gemeinsame Bundesausschuss hat die dort getroffenen Festlegungen zu Responseschwellen im Nachgang in seine Dossiervorlagen übernommen und hierzu ein Anhörungsverfahren durchgeführt [9,10].

Die methodischen Grundlagen für die vorgenommene Festlegung wurden bislang ausschließlich im Rahmen der Dokumentation und Würdigung der Anhörung zum damaligen Methodenpapier publiziert. Weitere Festlegungen für spezifische Instrumente sind zudem in einzelnen Dossierbewertungen des IQWiG zur Bewertung des Zusatznutzens von neuen Arzneimitteln beschrieben [11,12].

Das vorliegende Arbeitspapier bringt die an den verschiedenen Stellen publizierten Aspekte in einem gemeinsamen Dokument zusammen und umfasst zusätzlich eine ausführlichere Darstellung der Methoden und Ergebnisse als Ausgangsbasis für die damalige Festlegung. Das vorliegende Arbeitspapier dokumentiert das damalige Vorgehen und umfasst keine Aktualisierung der zugrunde liegenden Literatur.

4 Methoden

Nachfolgend wird das methodische Vorgehen für die im Jahr 2020 mit den Allgemeinen Methoden Version 6.0 [1] publizierte 15 %-Schwelle beschrieben.

Informationsbeschaffung

Es wurde eine fokussierte Informationsbeschaffung nach systematischen Übersichtsarbeiten zu MIDs patientenberichteter Endpunkte (PROs) in einer bibliografischen Datenbank (MEDLINE) durchgeführt (Suchdatum 22.05.2019). Die Übersichtsarbeiten wurden unter Berücksichtigung folgender Einschlusskriterien selektiert:

- Übersichtsarbeiten ab 2016
- Die Übersichtsarbeiten sollten Arbeiten einschließen, in denen MIDs empirisch ermittelt wurden.
- Die Übersichtsarbeiten sollten Arbeiten einschließen, in denen die MIDs für Instrumente ermittelt wurden, die patientenberichtet sind, mehr als 1 Item umfassen und deren Ausprägungen nicht direkt interpretierbar sind (abstrakte Skalen in Abgrenzung zu bspw. natürlichen Skalen, deren Ergebnis direkt interpretierbar ist).

Selektion der systematischen Übersichtsarbeiten

Die eingeschlossenen systematischen Übersichtsarbeiten sollten einen möglichst breiten Überblick über MIDs zu PROs in verschiedenen Indikationen und Populationen ermöglichen. Systematische Übersichtsarbeiten wurden nicht in die Untersuchung eingeschlossen, wenn die Studien zum großen Teil mit denen der bereits eingeschlossenen Übersichtsarbeiten überlappen.

Extraktion der MIDs

Die Volltexte der in die systematischen Übersichtsarbeiten eingeschlossenen Referenzen wurden beschafft. Die Publikationen wurden hinsichtlich der Methodik zur Ermittlung der jeweiligen MID gesichtet. Die Methodik zur Ermittlung der MID musste folgende Kriterien erfüllen:

- longitudinale Studie
- ankerbasierte MID
- patientenberichteter Anker (oder bei Kindern / Demenzkranken durch die Eltern bzw. Proxy)
- Global-Rating-of-Change-Anker
- Der Cut-off für den Global-Rating-of-Change-Anker sollte bei minimal, small, little oder höchstens moderate liegen, um die Ermittlung einer MID zu gewährleisten.

- Es musste sich um abstrakte PRO-Multi-Item-Skalen handeln.

Nur wenn die in den Publikationen dargestellte Methodik die oben genannten Kriterien erfüllten, wurden die Angaben zu den empirisch ermittelten MIDs extrahiert. Zudem wurde die jeweilige Skalenspannweite des Instruments identifiziert und die MID hierzu ins Verhältnis gesetzt.

5 Ergebnisse

5.1 Ergebnis der Informationsbeschaffung

Abbildung 1 zeigt das Ergebnis der fokussierten Informationsbeschaffung und der Studieneinklassifizierung gemäß den Kriterien für den Studieneinschluss. Die Suchstrategie für die Suche in bibliografischen Datenbanken finden sich in Anhang B. Die Suche fand am 21.5.2019 statt.

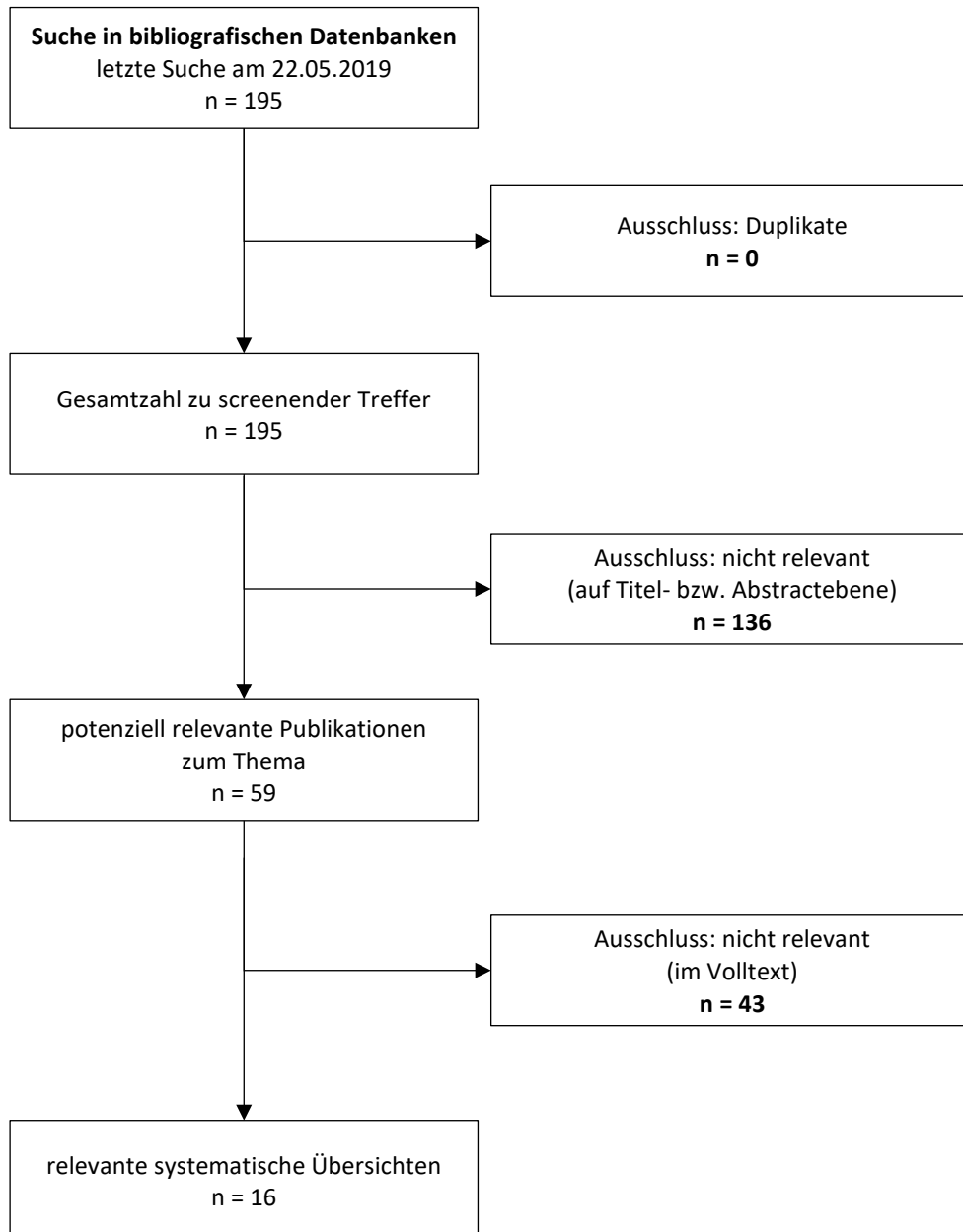


Abbildung 1: Ergebnis der fokussierten Informationsbeschaffung nach systematischen Übersichtsarbeiten zu MIDs

Durch die Informationsbeschaffung wurden 16 Übersichtsarbeiten identifiziert [13-28], von denen 8 [13,19,20,22,23,25,27,28] in die Untersuchungen eingeschlossen wurden. Diese schlossen Studien zu orthopädischen [22,28], pädiatrischen [20] und onkologischen [27] Indikationen oder zu einzelnen Erkrankungen wie Rheuma [19] oder chronisch obstruktiver Lungenerkrankung (COPD) [13] ein. Eine der Übersichtsarbeiten betrachtete ausschließlich Studien zur Ermittlung der MID von Instrumenten zu Erhebung von Fatigue [25], eine andere Fragebogen zu Erhebung der gesundheitsbezogenen Lebensqualität [23].

Auf die Extraktion der übrigen 8 Übersichtsarbeiten [14-18,21,24,26] wurde verzichtet, da diese ausschließlich orthopädische Indikationen abbilden und die Studien zum großen Teil mit denen der bereits eingeschlossenen Übersichtsarbeiten überlappen. Somit waren keine über die bereits extrahierten Übersichtsarbeiten hinausgehenden Informationen zu erwarten.

5.2 Ergebnisse zu MIDs aus den in die systematischen Übersichtsarbeiten eingeschlossenen Studien

Nachfolgend werden zunächst die extrahierten Übersichtsarbeiten kurz charakterisiert.

- Doganay 2016 [19] (Rheuma): Ziel der Studie war die Evaluation von MIDs in der Rheumatologie der vergangenen 2 Dekaden. Unter Berücksichtigung der Einschlusskriterien wurden zu 17 Instrumenten MIDs aus insgesamt 15 Studien extrahiert. Die Spanne der MIDs reicht von < 1 % bis 38 % der Skalenspannweite des jeweiligen Instruments. Weitere 21 der dort identifizierten Studien erfüllten nicht die Einschlusskriterien.
- St-Pierre 2016 [28] und Hao 2019 [22] (Orthopädie): In der Übersichtsarbeit St-Pierre wurden die psychometrischen Eigenschaften von PROs zur Evaluation von Symptomen und Funktionseinschränkungen bei Patienten mit Verletzungen der Rotatorenmanschette untersucht. Die Angaben aus 8 von 11 Studien zu 8 PRO-Instrumenten erfüllten die Einschlusskriterien für die vorliegende Untersuchung. Die Spanne der MIDs reicht von 6 % bis 23 %. Ziel der Übersichtsarbeit Hao 2019 war es, ankerbasierte MIDs für PROs bei Patientinnen und Patienten mit Schulterbeschwerden zu ermitteln. Insgesamt erfüllten 11 von 22 Studien die Einschlusskriterien und es wurden MIDs zu 9 verschiedenen Erhebungsinstrumenten extrahiert. Die Spannweite der MIDs über alle Instrumente liegt bei 4 % bis 23 % der Skalenspannweite.
- Ebrahim 2017 [20] (Pädiatrie): Die Übersichtsarbeit hat ankerbasierte MIDs für PROs in pädiatrischen Studien unterschiedlicher Erkrankungen untersucht. Für 10 Erhebungsinstrumente wurden die Angaben zu MIDs aus 7 eingeschlossenen Studien extrahiert. Die Spanne der MIDs reicht von 2 % bis 34 % der Skalenspannweite. Weitere 25 der dort identifizierten Studien erfüllten nicht die Einschlusskriterien.

- Nordin 2016 [25] (Fatigue): Ziel der Übersichtsarbeit war die Zusammenstellung und Bewertung publizierter MIDs für Pros zu Fatigue. Für 5 PROs wurden die Angaben zu MIDs aus 4 eingeschlossenen Studien berücksichtigt. Die Spanne der MIDs reicht von 2 % bis 24 % der Skalenspannweite. Weitere 37 der dort identifizierten Studien erfüllten nicht die Einschlusskriterien.
- Jayadevappa 2017 [23] (gesundheitsbezogene Lebensqualität): Ziel der Übersichtsarbeit war die Bewertung von MIDs bei der Messung gesundheitsbezogener Lebensqualität mit generischen und krankheitsspezifischen Instrumenten. Dafür wurden Instrumente in unterschiedlichen Indikationsgebieten untersucht. Für 24 Instrumente wurden die MIDs aus 23 Studien extrahiert. Die Spanne der MIDs reicht von < 1 % bis 38 % der Skalenspannweite. Weitere 57 der dort identifizierten Studien erfüllten nicht die Einschlusskriterien.
- Alma 2018 [13] (COPD): Für die Übersichtsarbeit wurde zunächst die verfügbare Evidenz zu MIDs von PRO-Instrumenten in der Indikation COPD identifiziert. Ziel war, für jedes PRO-Instrument die durchschnittliche MID auf Basis anker- und verteilungsbasierter MIDs durch Methoden Triangulation zu ermitteln. Für die vorliegende Arbeit wurden die MIDs zu 4 Instrumenten aus insgesamt 5 Studien extrahiert. Die Spanne reicht von 3 % bis 17 % der Skalenspannweite. Weitere 16 der dort identifizierten Studien erfüllten nicht die Einschlusskriterien.
- Ousmen 2018 [27] (Onkologie): Ziel der Übersichtsarbeit war die Prüfung der verwendeten statistischen Methoden zur MID-Bestimmung patientenberichteter Endpunkte in der Onkologie. Für 2 Instrumente wurden die MIDs aus 2 Studien herangezogen. Die Spanne reicht von < 1 % bis 27 % der Skalenspannweite. Weitere 44 der dort identifizierten Studien erfüllten nicht die Einschlusskriterien.

In die systematischen Übersichtsarbeiten wurden zwischen 21 [13] und 80 [23] Studien zu MIDs eingeschlossen. Ein Großteil dieser Studien erfüllte die Einschlusskriterien für die vorliegende Arbeit nicht. Der häufigsten Ausschlussgrund war, dass für die Ableitung der MID kein Global-Rating-of-Change-Anker herangezogen wurde. Weitere Ausschlussgründe waren u. a.: Die Methodik der MID-Ermittlung war unklar beschrieben, der Anker für die MID war nicht patientenberichtet, der Cut-off für den Global-Rating-of-Change-Anker war größer als minimal, die MID wurde auf Grundlage einer verteilungsbasierten Methodik ermittelt oder es handelte sich um Querschnittsstudien.

Die folgende Tabelle 1 zeigt die Verteilung der extrahierten MIDs im Verhältnis zur Skalen-Spannweite in %. Eine detailliertere Übersicht bezogen auf die einzelnen Übersichtsarbeiten und Instrumente findet sich in Anhang B.

Tabelle 1: Spannweite der empirisch erhobenen MIDs im Verhältnis zur Spannweite der jeweils verwendeten Skala in %

Übersichtsarbeit	Therapiegebiet	Spannweite der erhobenen MID im Verhältnis zur Spannweite der Skala in %
Doganay 2016 [19]	Rheuma	< 1–38
St-Pierre 2016 [28]	Orthopädie	6–23
Hao 2019 [22]	Orthopädie	4–23
Ebrahim 2017 [20]	Pädiatrie	2–34
Nordin 2016 [25]	Fatigue	2–24
Jayadevappa 2017 [23]	▪ Onkologie ▪ kardiovaskuläre Erkrankungen ▪ respiratorische Erkrankungen ▪ Erkrankungen des Urogenitaltrakts ▪ muskuloskelettale Erkrankungen ▪ Pädiatrie ▪ verschiedenes	3–18 9–16 3–17 6–18 1–23 < 1–38 < 1–23
Alma 2018 [13]	COPD	3–17
Ousmen 2018 [27]	Onkologie	< 1–27
COPD: chronisch obstruktive Lungenerkrankung; MID: Minimal important Difference		

Die Betrachtung von empirisch erhobenen MIDs zeigt über alle Indikationen hinweg eine große Variabilität der Ergebnisse von Studien zur Ermittlung von MIDs auf.

Die Werte der MIDs befinden sich zwischen < 1 % und 38 % der Skalenspannweite (siehe Tabelle 3 bis Tabelle 10 in Anhang B). Dabei ist hervorzuheben, dass in diese Aufstellung bereits nur solche Studien eingingen, die die festgelegten Mindestqualitätskriterien erfüllten (z. B. patientenberichteter Anker, Längsschnittstudie, siehe oben). Diese Kriterien stellen einen wesentlichen Bestandteil des von der Gruppe der McMaster-Universität vorgeschlagenen Instruments zur Beurteilung der Qualität von MIDs dar [5].

Trotz des Einhaltens der festgelegten Mindestqualitätskriterien zeigt die Methodik zur Ermittlung der MIDs zwischen den Studien erhebliche Abweichungen. Die MIDs wurden in den eingeschlossenen Studien überwiegend entweder als mittlere Veränderung (mean change) im Score derjenigen Patientinnen und Patienten die sich (minimal) spürbar im Vergleich zu Studienbeginn verändert hatten oder mithilfe von Receiver Operating Characteristic (ROC) Kurven bestimmt. Zum Teil wurden MIDs separat für Verbesserung und Verschlechterung, für unterschiedliche Patientengruppen, für verschiedene Recall-Zeiten sowie für verschiedene Domänen oder den Gesamtscore eines Instruments bestimmt.

Die Variabilität der bei den verschiedenen Absätzen abgeleiteten MIDs wird im Folgenden exemplarisch für den Western Ontario and McMaster Osteoarthritis Index (WOMAC)

beschrieben. Der WOMAC besteht aus 24 Items, die den Domänen Schmerz, Steifheit und körperliche Funktion zugeordnet werden. Es werden Scores für die einzelnen Domänen und ein Gesamtscore berechnet. In der Arbeit von Doganay 2016 [19] wurden in 11 Studien MIDs für den WOMAC berichtet. Von diesen erfüllten 8 Studien die Einschlusskriterien für die vorliegende Untersuchung. Aus diesen resultiert eine Vielzahl unterschiedlicher MIDs. Diese sind in der folgenden Tabelle 2 dargestellt.

Tabelle 2: Zusammenfassung der MIDs für den WOMAC aus 8 in die Übersichtsarbeit Doganay eingeschlossenen Studien [19]

Instrument	Domäne	MIDs in % der Skalenspannweite	% Spannweite MIDs
Western Ontario and McMaster Osteoarthritis Index (WOMAC) ^a	Schmerz	6,4; 7,5; 8,3; 11; 14,7; 17,5; 22,9; 23,3; 29,3; 33,1; 35,8	6,4–35,8
	Steifheit	2,9; 5,1; 6,3; 7,2; 10,1; 14,5; 18,8; 25,9; 33,2	2,9–33,2
	Körperliche Funktion	5,9; 6,7; 8; 8,1; 9,4; 10,3; 13,3; 19; 21,8; 25,9; 26,5; 30,5	5,9–30,5
	Gesamt score	1,6; 4; 6,7; 6,8; 8,2; 8,8; 9,6; 11,5; 12; 12,9; 20,5; 23,5; 28,1; 29,9	1,6–29,9
a: Skalenspannweite des Instruments: 0–10 Punkte			
MID: Minimal important Difference; WOMAC: Western Ontario and McMaster Osteoarthritis Index			

In 5 dieser Studien wurden die MIDs für den WOMAC als mittlere Veränderung im Vergleich zu Studienbeginn und in 2 Studien mithilfe von ROC Kurven bestimmt, in einer weiteren Studie wurden die MIDs auf Basis beider Methoden ermittelt und berichtet. In 2 Studien wurden MIDs für Verbesserung und Verschlechterung jeweils für die Domänen und für den Gesamtscore berechnet. In den Studien wurden die MIDs zu unterschiedlichen Recall-Zeiten (zwischen 2 Monaten und 24 Monate seit Studienbeginn) ermittelt. Zum Teil wurden MIDs jeweils für die einzelnen Domänen und den Gesamtscore ermittelt. Für 1 Studie wurde lediglich die MID für den Gesamtscore berichtet, in 2 weiteren Studien nur die MIDs für einzelne Domänen, nicht jedoch für den Gesamtscore.

Diese Ergebnisse zeigen noch einmal anschaulich die eingangs beschriebene Situation, dass sich für einzelne Instrumente keine einheitlichen MIDs identifizieren lassen. Obwohl in der vorliegenden Untersuchung durch die Einschlusskriterien bereits methodische Mindeststandards angelegt wurden, unterscheiden sich die Werte deutlich. Hinzu kommen Unterschiede durch die Patientengruppe, die Erkrankung, dem Schweregrad dieser Erkrankung oder auch die Richtung der Veränderung.

In der Gesamtschau aller Ergebnisse erscheinen die Werte am unteren Rand der erhobenen MIDs über alle Studien unrealistisch klein und bezüglich der Abgrenzung von spürbarer Veränderung und Messunsicherheit nicht plausibel. Die Frage, ob die Werte am oberen Rand

des Spektrums verlässlicher sind, lässt sich aufgrund fehlender akzeptierter Standards zur Qualitätsbewertung, wie auch fehlender Berichtsqualität von MID-Studien, nicht beantworten (siehe unten). Jedoch zeigt die systematische Betrachtung der extrahierten MIDs, dass der wesentliche Anteil empirisch ermittelter MIDs oberhalb der in der Literatur diskutierten generischen Schwelle von 10 % aber zumeist unterhalb von 20 % der jeweiligen Skalenspannweite liegt. Auf Basis der Ergebnisse dieser systematischen Erhebung von MIDs wurde somit ein Schwellenwert von 15 % als plausibel geeignet angesehen, hinreichend sicher eine für Patientinnen und Patienten spürbare Veränderung abzubilden. Diese Responseschwelle soll eine eher kleine Veränderung darstellen, der Wert kann aber durchaus oberhalb einer minimalen Schwelle liegen. Das heißt, bei der Festlegung dieser Schwelle geht es explizit nicht darum, ein Responsekriterium im Sinne einer individuellen MID zu identifizieren. Bei der Festlegung der 15 %-Schwelle handelt es sich um eine pragmatische Setzung vor dem Hintergrund der in den systematischen Übersichten identifizierten MIDs und nicht um einen anhand der Ergebnisse berechneten Wert. Die Methodik der MID-Identifizierung, zugehörige Qualitätskriterien und damit verbundene Ansätze zur Auswertung patientenberichteter Instrumente sind derzeit Gegenstand wissenschaftlicher Diskussionen. Die vorgenommene Setzung soll in dieser Situation eine pragmatische Lösung darstellen, die insbesondere im Rahmen der frühen Nutzenbewertung Klarheit für alle Beteiligten schafft.

6 Diskussion

Bei der vorgenommenen Festlegung einer 15 %-Schwelle handelt es sich um eine pragmatische Setzung, die vor dem Hintergrund des Fehlens einheitlicher instrumentenspezifischer Responseschwellen im Sinne einer individuellen MID sowie der beschriebenen methodischen Mängel bisher identifizierter MIDs vorgenommen wurde. Zudem fehlen in Publikationen zu MIDs häufig ausreichende Angaben, um deren Qualität beurteilen zu können [4], sodass es nicht möglich ist, aus der Vielzahl der vorliegenden MIDs für ein einzelnes Instrument den in der vorliegenden Situation verlässlichsten Wert zu identifizieren.

Im Methodenpapier des IQWiG wurde weiterhin festgelegt, dass Auswertungen mit einer Responseschwelle von 15 % der Skalenspannweite auch dann akzeptiert werden, wenn sie nicht im Studienprotokoll und Analyse plan festgelegt wurden (post-hoc Analyse der jeweiligen Studie). Da diese Auswertung für die Nutzenbewertung des IQWiG angefordert (prä-definiert) wird, wird das Risiko einer ergebnisgesteuerten Berichterstattung, das ansonsten bei post-hoc festgelegten Responsekriterien besteht, verringert. Durch diesen Schritt wurde für alle Beteiligten Verlässlichkeit insbesondere für das Verfahren der frühen Nutzenbewertung von Arzneimitteln geschaffen, welche Auswertungen akzeptiert werden.

6.1 Anhörung zum Verfahrensvorschlag

Die neue Setzung wurde erstmals mit dem Entwurf der Allgemeinen Methoden Version 6.0 Ende 2019 veröffentlicht, im anschließenden Anhörungsverfahren mit der Fachwelt diskutiert und im Jahr 2020 mit der finalen Version 6.0 etabliert [1,8]. Im Jahr 2021 hat der G-BA die neue Schwelle in die Dossievorlagen für die Bewertung des Zusatznutzens von neuen Arzneimitteln übernommen. Auch der G-BA hat hierzu ein Stellungnahmeverfahren einschließlich mündlicher Anhörung durchgeführt [9,10,29]. In den Diskussionen wurden insbesondere folgende Aspekte thematisiert.

Generische Responseschwelle versus „MID“

In den Anhörungen wurde diskutiert, ob das Responsekriterium von 15 % der Skalenspannweite zu unspezifisch und zu hoch sei, um eine MID darzustellen. Allerdings ist hierzu anzumerken, dass die neue Schwelle – wie bereits beschrieben – keine Schwelle im Sinne einer individuellen MID darstellen soll. Die Schwelle soll vielmehr hinreichend sicher eine für Patientinnen und Patienten spürbare Veränderung abbilden (siehe oben). Dabei wird davon ausgegangen, dass der Wert durchaus oberhalb der jeweiligen MID liegen kann. Diese Eigenschaft berücksichtigt auch die bestehende Variabilität von MIDs und trägt dazu bei, dass das Responsekriterium in unterschiedlichen Konstellationen nicht zu häufig unter die MID fällt. Gleichzeitig ist der Wert klein genug, um eine eher geringe Änderung abzubilden.

In diesem Zusammenhang ist es richtig, dass die Definition der Responseschwelle von 15 % der Skalenspannweite dazu führt, dass einige der bisher verwendeten Responsekriterien nicht mehr berücksichtigt werden. Das galt z. B für die von den Stellungnehmenden angeführten Responsekriterien für den European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire – Core 30 (QLQ-C30) von 10 Punkten oder den St. George's Respiratory Questionnaire (SGRQ) von 4 Punkten (Skalenspannweite des Instruments 0 bis 100 Punkte).

Zur Beurteilung dieses Sachverhalts ist es aber notwendig, die Qualität der infrage stehenden Responsekriterien zu betrachten. In den von den Stellungnehmenden genannten Beispielen gehen die Responsekriterien in der Regel auf Studien zurück, die dem aktuellen Stand der wissenschaftlichen Erkenntnis zu Studien zur Bestimmung von MIDs nicht entsprechen.

Für die MID von 4 Punkten für den SGRQ wird primär eine narrative Übersicht aus dem Jahr 2005 zitiert [30], andere Arbeiten aus den Jahren 2015 und 2018 beschreiben MIDs zwischen 6 und 9 Punkten [13,31].

So wird das Responsekriterium für den EORTC QLQ-C30 primär mit der Arbeit von Osoba et al. [32] begründet. In dem dort herangezogenen Datensatz erfüllte z. B. die Korrelation von Anker und Skala nicht die jetzt diskutierten Anforderungen. In dieser Situation hat die European Organisation for Research and Treatment of Cancer (EORTC), die für das Instrument verantwortlich zeichnet, ein Projekt zur Bestimmung von MIDs für den EORTC QLQ-C30 - Fragebogen begonnen [33]. Inzwischen liegen erste Ergebnisse zu MIDs für Gruppenunterschiede vor, Ergebnisse zu MIDs auf individueller Patientenebene stehen noch aus. Bereits in der vorliegenden Arbeit diskutieren die Autorinnen und Autoren, dass keine einheitliche MID für Gruppenunterschiede etabliert werden kann, sondern dass diese sich zwischen den einzelnen Skalen sowie verschiedenen Krankheitssettings unterscheiden [34] (zum aktuellen Umgang mit individuellen Schwellenwerten für den EORTC QLQ-C30 und seinen Zusatzmodulen siehe Anhang C.1).

Robuste Effektschätzung

Es wurde des Weiteren vonseiten insbesondere der stellungnehmenden pharmazeutischen Unternehmen die Befürchtung geäußert, dass durch das vorgeschlagene Vorgehen die Anforderungen erhöht werden und der Nachweis eines Zusatznutzens von Interventionen erschwert würde. Die Stellungnehmenden führten nicht aus, warum sie davon ausgehen, dass das vorgeschlagene Responsekriterium den Nachweis eines Zusatznutzens erschweren sollte, und legten weder im Stellungnahmeverfahren des IQWiG noch des G-BA Daten vor, die diese Annahme stützen würden. Generell ist nicht davon auszugehen, dass (robuste) Effekte nicht mehr nachweisbar sind, wenn sich alternativ eingesetzte Responsekriterien im unteren

Bereich der Spannweite einer Skala bewegen. Diese Einschätzung wurde durch die Diskussion in der wissenschaftlichen Erörterung zum IQWiG-Methodenpapier bestätigt.

Inhaltliche Bewertung bei natürlichen Skalen

Ein weiterer Diskussionspunkt betraf die Anwendbarkeit der Schwelle auf natürliche Skalen, die keine Spannweite haben, wie beispielsweise die 6-Minuten-Gehstrecke. Es ist richtig, dass das vorgeschlagene Responsekriterium von mindestens 15 % der Skalenspannweite z. B. für nach oben offene Skalen nicht herangezogen werden kann. Eine solche Anwendung ist aber auch nicht geplant. Die Festlegung adressiert komplexe Skalen (insbesondere psychometrische Skalen), deren Ergebnisse nicht ohne Weiteres interpretiert werden können. Das vorgeschlagene Responsekriterium 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. Da komplexe psychometrische Skalen stets über einen Maximalwert verfügen, kann das von den Stellungnehmenden angeführte Problem nicht auftreten (zum Umgang mit Sonderfällen siehe bspw. zum SF-36 in Anhang C.2).

Bewertung von „MIDs“

Generell wurden in den Stellungnahmen wenig konkrete Vorschläge gemacht, wie mit der Variabilität von MIDs und den bestehenden Schwierigkeiten bei der Bewertung aktuell vorliegender Studien zu MIDs umgegangen werden kann. Von Devji et al. wurde 2020 ein Instrument zur Beurteilung der Aussagekraft ankerbasierter Festlegungen von MIDs veröffentlicht [5]. In dieser Arbeit beschreibt eine Gruppe um Wissenschaftlerinnen und Wissenschaftler der McMaster-Universität die Entwicklung von Qualitätskriterien für Studien zur Ermittlung von MIDs und die Ergebnisse einer Studie zur Reliabilität des neu entwickelten Instruments. Grundsätzlich fokussieren sie dabei allein auf ankerbasierte Verfahren; verteilungsbasierte Verfahren werden per se als ineffektiv und als die Patientenperspektive vernachlässigend eingeschätzt. Die Qualitätskriterien, die die Autorinnen und Autoren der Arbeit für die Beurteilung der Aussagekraft einer MID benennen, sind vor dem Hintergrund der Diskussion der letzten Jahre nachvollziehbar. Das Instrument besteht aus 5 Kernkriterien (Beantwortung des Fragebogens, für den die MID bestimmt werden soll, und des Ankers durch Patientinnen und Patienten (bzw. wenn notwendig eine Vertretung); leichte Verständlichkeit und Relevanz des Ankers für Patientinnen und Patienten; eine gute Korrelation zwischen Fragebogen und Anker; Präzision der MID; der Schwellenwert des Ankers für die Bestimmung der MID zeigt einen kleinen, aber relevanten Unterschied an) und 4 zusätzlichen Kriterien für MIDs, die mithilfe von Abfragen zu Änderungen des Ankers (Transition Ratings) erfolgen (optimale Zeitspanne zwischen den Messungen; zufriedenstellende Korrelation zwischen Fragebogen und Anker [Transition Item] bei der Follow-up-Messung; Korrelation des Ankers [Transition Item] mit dem Ausgangswert des Fragebogens; höhere Korrelation zwischen Änderung des Ankers [Transition Item] und Änderung des Fragebogens als zwischen Änderung

des Ankers und Wert des Fragebogens zum Follow-up). Diese ergänzenden Kriterien sind relevant, weil ein großer Teil von MIDs über die Abfrage der Änderungen von Fragebogen und Anker erhoben wird.

Mangelnde Publikationsqualität von „MID-Studien“

Interessant für die Diskussion um den Methodenvorschlag des IQWiG sind die Ergebnisse der Reliabilitätsstudie [5]. Für die Untersuchung der Reliabilität des neu entwickelten Instruments zur Beurteilung der Aussagekraft ankerbasierter Festlegungen von MIDs nutzte die Gruppe eine Zufallsstichprobe aus einer Datenbank empirisch erhobener MIDs. In dieser Stichprobe konnten die Autorinnen und Autoren die Reliabilität der Kriterien für MID-Ermittlungen über Abfragen zur Änderung des Ankers nicht beurteilen, weil die dafür notwendigen Angaben nur in einem sehr geringen Teil der Studien zur Ermittlung der MIDs berichtet wurden. Das Instrument wurde darüber hinaus von der gleichen Arbeitsgruppe in einer systematischen Übersichtsarbeit zu MIDs zu PROs eingesetzt. Auch hier zeigte sich, dass wesentliche Qualitätskriterien für MIDs in den eingeschlossenen Studien nicht berichtet wurden [4]. Diese Daten decken sich mit Erfahrungen aus Dossierbewertungen zur Bewertung des Zusatznutzens von neuen Arzneimitteln, in denen ebenfalls häufig nicht alle notwendigen Angaben zur Prüfung der Aussagekraft einer MID vorliegen. Der sofortige Einsatz dieses neuen Instruments in der Dossierbewertung würde deshalb voraussichtlich wegen fehlender Angaben zu den Qualitätskriterien zu einem hohen Anteil negativer Bewertungen der Validität der MIDs führen, mit der Konsequenz das die vorgelegten Analysen nicht für die Bewertung herangezogen werden. Der Vorschlag der generischen Responseschwelle mit Bezug zur Skalenspannweite erscheint auch unter diesem Gesichtspunkt als ein sinnvoller Schritt, zumindest für den Zeitraum, bis sich die Qualität der Studien zur Ermittlung von MIDs verbessert hat.

6.2 Internationale Rezeption des 15 %-Ansatzes

Zwischenzeitlich findet die Festlegung der generischen Responseschwelle für Responderanalysen durch das IQWiG auch international Aufmerksamkeit. Um Konsistenz in der Interpretation der Ergebnisse verschiedener PRO Instrumente zu gewährleisten und wegen des Fehlens publizierter MIDs wurden in der Studie von Rea et al. Auswertungen auf Basis der Responseschwelle von 15 % der Skalenspannweite durchgeführt [35]. Die Initiative Common Sense Oncology und die EORTC empfehlen wegen des Fehlens evidenzbasierter Schwellenwerte einen pragmatischen Ansatz wie die Setzung des 15 % Responsekriteriums. Um Unsicherheiten bezüglich der generellen Anwendbarkeit dieses Kriteriums zu adressieren, empfehlen die Autoren Sensitivitätsanalysen [36].

7 Fazit

Für die Auswertung mittels komplexer Skalen erhobener patientenberichteter Endpunkte werden häufig Responderanalysen auf Basis eines Responsekriterium im Sinne einer individuellen MID durchgeführt. Unter einer MID wird dabei der kleinste Unterschied im Score verstanden, der für einen Patienten oder einer Patientin spürbar ist und als Verbesserung oder Verschlechterung bewertet wird. Allerdings ist die Bestimmung einer MID mit methodischen Problemen verbunden und für einzelne Instrumente wird häufig eine Vielzahl von MIDs publiziert.

Ziel der Untersuchung war es daher, eine generische Responseschwelle für Responderanalysen bei komplexen Skalen, insbesondere solchen zur Erhebung patientenberichteter Endpunkte, zu identifizieren, die hinreichend sicher eine für Patientinnen und Patienten spürbare Veränderung darstellt.

- Die Sichtung von systematischen Übersichtsarbeiten zur Identifikation von MIDs für PROs in verschiedenen Indikationen bestätigte die oben genannte Problematik:
 - Bei der Bestimmung von MIDs werden häufig wesentliche methodische Qualitätskriterien nicht eingehalten. So konnte ein Großteil der in die systematischen Übersichtsarbeiten eingeschlossenen Studien nicht für die Extraktion der MIDs herangezogen werden.
 - Auch die in methodisch sachgerecht durchgeführte Studien empirisch erhobenen MIDs weisen über alle PRO Instrumente hinweg eine große Variabilität auf.
- Jedoch zeigt die systematische Betrachtung der extrahierten MIDs, dass der wesentliche Anteil empirisch ermittelter MIDs oberhalb der in der Literatur diskutierten generischen Schwelle von 10 %, aber zumeist unterhalb von 20 % der jeweiligen Skalenspannweite liegt. Auf Basis dieser Ergebnisse wurde somit ein Schwellenwert von 15 % der Skalenspannweite als geeignet angesehen, hinreichend sicher eine für Patientinnen und Patienten spürbare Veränderung abzubilden.
- Hierbei handelt es sich um eine pragmatische Setzung vor dem Hintergrund der in den systematischen Übersichten identifizierten MIDs und nicht um einen anhand der numerischen Ergebnisse errechneten Wert. Bei dieser Setzung handelt es sich explizit nicht um ein Responsekriterium im Sinne einer individuellen MID sondern um eine generisch anwendbare Responseschwelle für PRO-Instrumente.
- In den Allgemeinen Methoden Version 6 [1] wurde weiterhin festgelegt, dass Auswertungen mit einer Responseschwelle von 15 % der Skalenspannweite auch dann akzeptiert werden, wenn sie post hoc (also nicht im Studienprotokoll der jeweiligen Studie prä-definiert) durchgeführt werden. Damit wurde für alle Beteiligten Verlässlichkeit insbesondere für das Verfahren der frühen Nutzenbewertung von Arzneimitteln geschaffen, welche Auswertungen akzeptiert werden.

8 Literatur

1. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden: Version 6.0 [online]. 2020 [Zugriff: 11.07.2023]. URL: https://www.iqwig.de/methoden/allgemeine-methoden_version-6-0_abgeloest.pdf.
2. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989; 10(4): 407-415. [https://doi.org/10.1016/0197-2456\(89\)90005-6](https://doi.org/10.1016/0197-2456(89)90005-6).
3. Schünemann HJ, Akl EA, Guyatt GH. Interpreting the results of patient reported outcome measures in clinical trials: the clinician's perspective. *Health Qual Life Outcomes* 2006; 4: 62. <https://doi.org/10.1186/1477-7525-4-62>.
4. Carrasco-Labra A, Devji T, Qasim A et al. Minimal important difference estimates for patient-reported outcomes: A systematic survey. *J Clin Epidemiol* 2021; 133: 61-71. <https://doi.org/10.1016/j.jclinepi.2020.11.024>.
5. Devji T, Carrasco-Labra A, Qasim A 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. <https://doi.org/10.1136/bmj.m1714>.
6. Ringash J, O'Sullivan B, Bezjak A et al. Interpreting clinically significant changes in patient-reported outcomes. *Cancer* 2007; 110(1): 196-202. <https://doi.org/10.1002/cncr.22799>.
7. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res* 2011; 11(2): 171-184. <https://doi.org/10.1586/erp.11.9>.
8. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Dokumentation und Würdigung der Anhörung zum Entwurf der Allgemeinen Methoden 6.0 [online]. 2020 [Zugriff: 11.07.2023]. URL: https://www.iqwig.de/methoden/allgemeine-methoden_dwa-entwurf-fuer-version-6-0_v1-0.pdf.
9. Gemeinsamer Bundesausschuss. Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Verfahrensordnung: Änderung der Modulvorlage in der Anlage II zum 5. Kapitel [online]. 2021 [Zugriff: 11.02.2026]. URL: https://www.g-ba.de/downloads/40-268-8139/2021-12-16_VerfO_Aenderung-Modulvorlage-Anlage-II-Kap-5_TrG.pdf.
10. Gemeinsamer Bundesausschuss. Zusammenfassende Dokumentation zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Verfahrensordnung: Änderung der Modulvorlage in der Anlage II zum 5. Kapitel [online]. 2021 [Zugriff: 11.02.2026]. URL: https://www.g-ba.de/downloads/40-268-8140/2021-12-16_VerfO_Aenderung-Modulvorlage-Anlage-II-Kap-5_ZD.pdf.

11. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Atezolizumab (hepatozelluläres Karzinom) – Nutzenbewertung gemäß § 35a SGB V; Dossierbewertung [online]. 2021 [Zugriff: 11.07.2023]. URL: https://www.iqwig.de/download/a20-97_atezolizumab_nutzenbewertung-35a-sgb-v_v1-0.pdf.
12. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Osimertinib (NSCLC, adjuvant) – Nutzenbewertung gemäß § 35a SGB V; Dossierbewertung [online]. 2021 [Zugriff: 11.07.2023]. URL: https://www.iqwig.de/download/a21-86_osimertinib_nutzenbewertung-35a-sgb-v_v1-0.pdf.
13. Alma H, de Jong C, Tsiligianni I et al. Clinically relevant differences in COPD health status: systematic review and triangulation. *European Respiratory Journal* 2018; 52(3): 09. <https://doi.org/10.1183/13993003.00412-2018>.
14. Celik D, Coban O, Kilicoglu O. Minimal Clinically Important Difference of Commonly Used Hip, Knee, Foot and Ankle Specific Questionnaires: A Systematic Review. *Journal of Clinical Epidemiology* 2019; 09: 09. <https://doi.org/10.1016/j.jclinepi.2019.04.017>.
15. Chung AS, Copay AG, Olmscheid N et al. Minimum Clinically Important Difference: Current Trends in the Spine Literature. *Spine* 2017; 42(14): 1096-1105. <https://doi.org/10.1097/BRS.0000000000001990>.
16. Copay AG, Chung AS, Eyberg B et al. Minimum Clinically Important Difference: Current Trends in the Orthopaedic Literature, Part I: Upper Extremity: A Systematic Review. *JBJS Reviews* 2018; 6(9): e1. <https://doi.org/10.2106/JBJS.RVW.17.00159>.
17. Copay AG, Eyberg B, Chung AS et al. Minimum Clinically Important Difference: Current Trends in the Orthopaedic Literature, Part II: Lower Extremity: A Systematic Review. *JBJS Reviews* 2018; 6(9): e2. <https://doi.org/10.2106/JBJS.RVW.17.00160>.
18. Devji T, Guyatt GH, Lytvyn L 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. <https://doi.org/10.1136/bmjopen-2016-015587>.
19. Doganay Erdogan B, Leung YY, Pohl C et al. Minimal Clinically Important Difference as Applied in Rheumatology: An OMERACT Rasch Working Group Systematic Review and Critique. *Journal of Rheumatology* 2016; 43(1): 194-202. <https://doi.org/10.3899/jrheum.141150>.
20. Ebrahim S, Vercammen K, Sivanand A et al. Minimally Important Differences in Patient or Proxy-Reported Outcome Studies Relevant to Children: A Systematic Review. *Pediatrics* 2017; 139(3). <https://doi.org/10.1542/peds.2016-0833>.

21. Goldsmith ES, Taylor BC, Greer N et al. Focused Evidence Review: Psychometric Properties of Patient-Reported Outcome Measures for Chronic Musculoskeletal Pain. *Journal of General Internal Medicine* 2018; 33(Suppl 1): 61-70. <https://doi.org/10.1007/s11606-018-4327-8>.
22. Hao Q, Devji T, Zeraatkar D 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. <https://doi.org/10.1136/bmjopen-2018-028777>.
23. Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life-a systematic review. *Journal of Clinical Epidemiology* 2017; 89: 188-198. <https://doi.org/10.1016/j.jclinepi.2017.06.009>.
24. MacKay C, Clements N, Wong R et al. A Systematic Review of Estimates of the Minimally Clinically Important Difference and Patient Acceptable Symptom State of the Western Ontario and McMaster Universities Osteoarthritis Index in Patients who Underwent Total Hip and Total Knee Replacement. *Osteoarthritis & Cartilage* 2019; 13: 13. <https://doi.org/10.1016/j.joca.2019.05.002>.
25. Nordin A, Taft C, Lundgren-Nilsson A et al. Minimal important differences for fatigue patient reported outcome measures-a systematic review. *BMC Medical Research Methodology* 2016; 16: 62. <https://doi.org/10.1186/s12874-016-0167-6>.
26. Nwachukwu BU, Runyon RS, Kahlenberg CA et al. How are we measuring clinically important outcome for operative treatments in sports medicine? *Physician & Sportsmedicine* 2017; 45(2): 159-164. <https://doi.org/10.1080/00913847.2017.1292108>.
27. Ousmen A, Touraine C, Deliu N et al. Distribution- and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review. *Health & Quality of Life Outcomes* 2018; 16(1): 228. <https://doi.org/10.1186/s12955-018-1055-z>.
28. St-Pierre C, Desmeules F, Dionne CE et al. Psychometric properties of self-reported questionnaires for the evaluation of symptoms and functional limitations in individuals with rotator cuff disorders: a systematic review. *Disability & Rehabilitation* 2016; 38(2): 103-122. <https://doi.org/10.3109/09638288.2015.1027004>.
29. Gemeinsamer Bundesausschuss. Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Verfahrensordnung: Änderung der Modulvorlage in der Anlage II zum 5. Kapitel vom 16. Dezember 2021 [online]. 2021 [Zugriff: 22.10.2025]. URL: https://www.g-ba.de/downloads/39-261-5217/2021-12-16_VerfO_Aenderung-Modulvorlage-Anlage-II-Kap-5.pdf.
30. Jones PW. St. George's Respiratory Questionnaire: MCID. *COPD* 2005; 2(1): 75-79. <https://doi.org/10.1081/copd-200050513>.

31. Welling JB, Hartman JE, Ten Hacken NH et al. The minimal important difference for the St George's Respiratory Questionnaire in patients with severe COPD. *Eur Respir J* 2015; 46(6): 1598-1604. <https://doi.org/10.1183/13993003.00535-2015>.
32. Osoba D, Rodrigues G, Myles J et al. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol* 1998; 16(1): 139-144. <https://doi.org/10.1200/JCO.1998.16.1.139>.
33. Musoro ZJ, Hamel JF, Ediebah DE 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. <https://doi.org/10.1136/bmjopen-2017-019117>.
34. Musoro JZ, Coens C, Sprangers MAG et al. Minimally important differences for interpreting EORTC QLQ-C30 change scores over time: A synthesis across 21 clinical trials involving nine different cancer types. *Eur J Cancer* 2023; 188: 171-182. <https://doi.org/10.1016/j.ejca.2023.04.027>.
35. Rea D, Boquimpani C, Mauro MJ et al. Health-related quality of life of patients with resistant/intolerant chronic phase chronic myeloid leukemia treated with asciminib or bosutinib in the phase 3 ASCEMBL trial. *Leukemia* 2023; 37(5): 1060-1067. <https://doi.org/10.1038/s41375-023-01888-y>.
36. Tannock IF, Pe ML, Booth CM et al. Importance of responder criteria for reporting health-related quality-of-life data in clinical trials for advanced cancer: recommendations of Common Sense Oncology and the European Organisation for Research and Treatment of Cancer. *Lancet Oncol* 2025; 26(9): e499-e507. [https://doi.org/10.1016/S1470-2045\(25\)00288-8](https://doi.org/10.1016/S1470-2045(25)00288-8).
37. Wong SS, 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.
38. Maruish ME. User's manual for the SF-36v2 Health Survey. Lincoln: QualityMetric; 2011.
39. Bjorner JB. Persönliche Kommunikation zu "Score Range for the SF-36v2 and SF-12v2 Health Surveys" [unveröffentlicht]. 2022.
40. Bjorner JB. Score Range for the SF-36v2 & SF-12v2 Health Surveys; Standard & Acute Versions [online]. 2022 [Zugriff: 06.01.2026]. URL: <https://www.qualitymetric.com/wp-content/uploads/2022/03/QM-SF-36v2-and-SF-12v2-DataSheet-v2.pdf>.
41. HealthMeasures. PROMIS (Patient-Reported Outcomes Measurement Information System) [online]. 2023 [Zugriff: 09.02.2026]. URL: <https://www.healthmeasures.net/explore-measurement-systems/promis>.
42. Peipert J, Cella Dea. Persönliche Kommunikation zu "Scale range of PROMIS measures" [unveröffentlicht]. 2023.

Anhang A Bibliografische Recherche

1. MEDLINE

Suchoberfläche: Ovid

- Ovid MEDLINE(R) ALL 1946 to May 21, 2019

Es wurden folgende Filter übernommen:

- Systematische Übersicht: Wong [37]

#	Searches
1	minimal clinically important difference/
2	(minim* adj2 important* adj1 (change* or improvement* or difference*)).ti,ab.
3	(minim* adj1 clinical* adj2 (change* or improvement* or difference*)).ti,ab.
4	((clinical* or minim*) adj1 (meaningful* or important* or significan*) adj1 (change* or improvement* or difference* or measure*)).ti.
5	or/1-4
6	cochrane database of systematic reviews.jn.
7	(search or medline or systematic review).tw.
8	meta analysis.pt.
9	or/6-8
10	9 not (exp animals/ not humans.sh.)
11	and/5,10
12	11 and (english or german).lg.
13	..l/ 12 yr=2016-Current

Anhang B Tabellen zu den identifizierten MIDs in den einzelnen systematischen Übersichtsarbeiten

Tabelle 3: Zusammenfassung der MIDs der aus der Übersichtsarbeit Doganay berücksichtigten Studien [19]

Instrument	Spannweite der Skala	% Spannweite MIDs ^a (Anzahl Studien)
Western Ontario and McMaster Osteoarthritis Index (WOMAC)	0–10	2–36 % (8)
Hospital for Special Surgery (HSS) Total Elbow Scoring Systems	0–100	18 (1)
Hospital for Special Surgery (HSS) short form	0–100	20 (1)
Mayo Clinic Mayo Performance Index for the Elbow	0–100	15 (1)
Elbow Function Assessment Scale (EFA)	0–100	17 (1)
Manchester-Oxford foot questionnaire (MOXFQ)	0–100	14–25 (1)
Health Assessment Questionnaire Disability Index (HAQ-DI)	0–3	3–13 (2)
International Knee Documentation Committee Subjective Knee Form (IKDC SKF)	0–100	6–17 (1)
Cincinnati Knee Rating System (CKRS)	0–100	14–16 (1)
Lower Extremity Functional Scale (LEFS)	0–100	< 1–13 (1)
Knee Outcome Survey-Activities of Daily Living Scale (ADLS)	0–100	2–11 (1)
Disabilities of the Arm, Shoulder, and Hand (DASH)	0–100	10 (1)
Quick Disabilities of the Arm, Shoulder, and Hand (Quick DASH)	0–100	14 (1)
Patient Rated Wrist Evaluation (PRWE)	0–100	14 (1)
Hip Dysfunction and Osteoarthritis Outcome Score (HOOS)	0–100	17–38 (1)
Intermittent and constant osteoarthritis pain (ICOAP)	0–100	18–19 (1)
Knee injury and Osteoarthritis Outcome Score Physical function Short-form (KOOS-PS)	0–100	2–8 (1)
a. prozentualer Anteil der jeweiligen MID an der Skalenspannweite des Instruments MID: Minimal important Difference		

Tabelle 4: Zusammenfassung der MIDs der aus der Übersichtsarbeit St. Pierre berücksichtigten Studien [28]

Instrument	Spannweite der Skala	% Spannweite MIDs^a (Anzahl Studien)
American Shoulder and Elbow Surgeon questionnaire (ASES),	0–100	6–17 (2)
Simple Shoulder Test (SST)	0–12	17–18 (2)
Disability of the Arm, Shoulder and Hand (DASH)	0–100	11–12 (2)
Quick Disability of the Arm, Shoulder and Hand (Quick DASH)	0–100	8–13 (3)
Shoulder Pain and Disability Index (SPADI)	0–100	8–23 (2)
Oxford Shoulder Score (OSS)	0–48	8–15 (2)
Penn Shoulder Score (PSS)	0–100	11 (1)
Western Ontario Rotator Cuff Index (WORC)	0–2100	11–17 (1)
a. prozentualer Anteil der jeweiligen MID an der Skalenspannweite des Instruments MID: Minimal important Difference		

Tabelle 5: Zusammenfassung der MIDs der aus der Übersichtsarbeit Hao berücksichtigten Studien [22]

Instrument	Spannweite der Skala	% Spannweite MIDs^a Anzahl Studien
American Shoulder and Elbow Surgeon questionnaire (ASES),	0–100	6–17 (3)
Constant Shoulder Score (CSS)	0–100	6–17 (3)
Disability of the Arm, Shoulder and Hand (DASH)	0–100	4–22 (5)
Quick Disability of the Arm, Shoulder and Hand (Quick DASH)	0–100	8–13 (2)
Michigan Hand Outcomes Questionnaire (MHQ)	0–100	17 (1)
Oxford Shoulder Score (OSS)	0–48	8–15 (4)
Shoulder Function Assessment Scale (SFA)	0–70	18 (1)
Shoulder Pain and Disability Index (SPADI)	0–100	15–23 (3)
Simple Shoulder Test (SST)	0–12	12–18 (3)
a. prozentualer Anteil der jeweiligen MID an der Skalenspannweite des Instruments MID: Minimal important Difference		

Tabelle 6: Zusammenfassung der MIDs der aus der Übersichtsarbeit Ebrahim berücksichtigten Studien [20]

Instrument	Spannweite der Skala	% Spannweite MID^a (Anzahl Studien)
Sino-Nasal Outcome Test (SNOT-22)	0–110	8 (1)
Aberdeen Varicose Veins Questionnaire (AVVQ)	0–60	4 (1)
Oxford hip score	12–60	18 (1)
Oxford knee score	12–60	8 (1)
Childhood Health Assessment Questionnaire (CHAQ)	0–3	6–8 (1)
Rhino-conjunctivitis Total Symptom Score (RTSS)	0–18	4–10 (1)
pain and functional dimensions of Western Ontario and McMaster Osteoarthritis Index (WOMAC)	0–100	21–34 (2)
Acne-specific Quality of Life Questionnaire (Acne- QoL) ^b	0–24 (0–30)	13–17 (1)
Knee injury and Osteoarthritis Outcome Score (KOOS)	0–100	2–20 (1)
Test for Respiratory and Asthma Control in Kids (TRACK)	0–100	8–10 (1)
a. prozentualer Anteil der jeweiligen MID an der Skalenspannweite des Instruments b. unterschiedliche Spannweiten für Subskalen MID: Minimal important Difference		

Tabelle 7: Zusammenfassung der MIDs der aus der Übersichtsarbeit Nordin berücksichtigten Studien [25]

Instrument	Spannweite der Skala	% Spannweite MID^a (Anzahl Studien)
Schwartz Cancer Fatigue Scale (SCFS)	6–30	9–24 (1)
Profile of Mood States-Fatigue (POMS-F)	0–28	8–20 (1)
Fatigue Associated with Depression Questionnaire (FAsD)	1–5	13–17 (1)
Quality of Life Inventory in Epilepsy (QOLIE-31)	0–100	2–5 (1)
FACIT-Fatigue	0–52	19 (1)
a. prozentualer Anteil der jeweiligen MID an der Skalenspannweite des Instruments MID: Minimal important Difference		

Tabelle 8: Zusammenfassung der MIDs der aus der Übersichtsarbeit Jayadevappa berücksichtigten Studien (mehreseitige Tabelle) [23]

Instrument	Spannweite der Skala	% Spannweite MIDs ^a (Anzahl Studien)
kardiovaskuläre Erkrankungen		
Barthel Index	0–20	9 (1)
Stroke Rehabilitation Assessment of Movement (STREAM) ^b	0–20 (0–30)	10–16 (1)
Erkrankungen des Urogenitaltrakts		
Incontinence Quality of Life (I-QOL)	0–100	6–7 (1)
Qualiveen-30	0–4	9–18 (1)
respiratorische Erkrankungen		
clinical COPD questionnaire (CCQ)	0–6	3–17 (1)
Visual Simplified Respiratory Questionnaire (VSRQ)	0–80	4 (1)
King's Brief Interstitial Lung Disease Questionnaire (K-BILD)	0–100	8–12 (1)
control of allergic rhinitis and asthma test (CARAT)	0–30	12 (1)
Pädiatrie		
Childhood Health Assessment Questionnaire (CHAQ) ^b	0–24 (0–3)	< 1–8 (1)
Acute Otitis Media Severity of Symptom (AOM-SOS)	0–10	38 (1)
Verschiedene		
Graves' ophthalmopathy quality of life questionnaire (GO-QOL)	0–100	5–8 (1)
Dermatology Life Quality Index (DLQI)	0–30	23 (1)
Health-Related Quality of Life for Eating Disorders questionnaire version-2 (HeRQoLEDv2)	0–100	< 1–15 (1)
Gastrointestinal Quality-of-Life Index (GIQLI) ^b	0–144 (0–100)	6–8 (1)
muskuloskelettale Erkrankungen		
Roland-Morris Disability Questionnaire (RMDQ)	0–24	21 (1)
Scoliosis Research Society-22 patient questionnaire (SRS-22) ^b	20–100(5–25)	1–16 (1)
Michigan Hand Outcomes Questionnaire (MHQ)	0–100	3–23 (1)
Quick Disabilities of the Arm, Shoulder, and Hand (Quick DASH)	0–100	13–14 (3)
Knee quality of life (KQoL-26)	0–100	3–18 (1)
Disability of the Arm, Shoulder and Hand (DASH)	0–100	10–11 (2)
Patient Rated Wrist Evaluation (PRWE)	0–100	14 (1)
Onkologie		
European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire – Core 30 (EORTC QLQ-C30)	0–100	7–18(1)
Edmonton Symptom Assessment System (ESAS) ^b	0–90 (0–20, 0–60)	3–12 (1)

Tabelle 8: Zusammenfassung der MIDs der aus der Übersichtsarbeit Jayadevappa berücksichtigten Studien (mehrseitige Tabelle) [23]

Instrument	Spannweite der Skala	% Spannweite MIDs^a (Anzahl Studien)
Immune Thrombocytopenic Purpura Patient Assessment Questionnaire (ITP-PAQ)	0–100	3–16 (1)
a. prozentualer Anteil der jeweiligen MID an der Skalenspannweite des Instruments b. unterschiedliche Spannweiten für Gesamtscore und Subskalen MID: Minimal important Difference		

Tabelle 9: Zusammenfassung der MIDs der aus der Übersichtsarbeit Alma berücksichtigten Studien [13]

Instrument	Spannweite der Skala	% Spannweite MIDs^a (Anzahl Studien)
Clinical COPD Questionnaire (CCQ)	0–6	3–17 (2)
COPD Assessment Test (CAT)	0–40	3–8 (3)
St. George's Respiratory Questionnaire (SGRQ)	0–100	6–13 (1)
Quality of life for respiratory illness questionnaire (QoLRIQ)	0–7	7 (1)
a. prozentualer Anteil der jeweiligen MID an der Skalenspannweite des Instruments MID: Minimal important Difference		

Tabelle 10: Zusammenfassung der MIDs der aus der Übersichtsarbeit Ousmen berücksichtigten Studien [27]

Instrument	Spannweite der Skala	% Spannweite MIDs^a (Anzahl Studien)
European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire – Core 30 (EORTC QLQ-C30)	0–100	< 1–27 (2)
European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire – Breast Cancer (EORTC QLQ-BR23)	0–100	< 1–20 (1)
a. prozentualer Anteil der jeweiligen MID an der Skalenspannweite des Instruments MID: Minimal important Difference		

Anhang C Spezifische Instrumente

Für die nachfolgenden Instrumente wurden im Rahmen von Dossierbewertungen Konkretisierungen zur Umsetzung des 15%-Ansatzes publiziert, die nachfolgend dargestellt sind.

C.1 EORTC QLQ-C30 und seine Zusatzmodule: Eignung der Responseschwellen von 10 Punkten für Auswertungen der EORTC-Module

Der EORTC QLQ-C30 erhebt anhand von 30 Items sowohl die gesundheitsbezogene Lebensqualität als auch generelle Symptome von Krebspatientinnen und -patienten. Das Instrument besteht aus einer Skala zum globalen Gesundheitsstatus und 5 Funktionsskalen sowie 8 Symptomskalen. Die Items werden auf einer Skala von 1 bis 4, für die beiden Fragen zum globalen Gesundheitsstatus von 1 bis 7, bewertet. Zur Auswertung werden die Skalen in Werte von 0 bis 100 transformiert.

Die verschiedenen Skalen umfassen jeweils unterschiedlich viele Items. Die Funktionsskalen umfassen die körperliche Funktion (5 Items), die Rollenfunktion (2 Items), die emotionale Funktion (4 Items), die kognitive Funktion (2 Items) und die soziale Funktion (2 Items). Die Skala zum globalen Gesundheitsstatus umfasst 2 Items. Die Symptomskalen umfassen Fatigue (3 Items), Schmerzen (2 Items), Übelkeit und Erbrechen (2 Items), Dyspnoe, Schlaflosigkeit, Appetitverlust, Verstopfung und Diarrhö (jeweils 1 Item).

Auch für die Skalen des EORTC-QLQ C30 hat sich, u. a. basierend auf der Arbeit von Osoba [32], international eine MID von 10 Punkten für alle Skalen etabliert, die jeweils 10 % der Skalenspannweite entspricht. Vor diesem Hintergrund wurde geprüft, ob basierend auf der Konstruktion und Item-Anzahl der jeweiligen Skalen, für Auswertungen mit einem Responsekriterium von 10 Punkten gegenüber einer Auswertung mit einem Responsekriterium von 15 % der Skalenspannweite der EORTC-Skalen (entspricht 15 Punkten) Unterschiede in den Ergebnissen entstehen.

Tabelle 11 gibt einen Überblick über die Spannweiten der rohen Werte und kleinstmöglichen Änderungsschritte der einzelnen Skalen in Abhängigkeit von der Item-Anzahl pro Skala. Aus dem kleinstmöglichen Änderungsschritt ergibt sich für jede Skala, ob Auswertungen der beiden Responsekriterien (10 Punkte / 15 Punkte) zu unterschiedlichen Ergebnissen führen.

Tabelle 11: Bedeutung der unterschiedlichen Responseschwellen 10 Punkte / 15 Punkte für die Skalen des EORTC QLQ-C30

EORTC QLQ-C30 Skala (Anzahl Items, Wertebereich)^a	Spannweite je Skala [rohe Werte]^b	Anzahl möglicher Änderungsschritte der Skala^c [n]	Kleinstmöglicher Änderungsschritt auf normierter Skala (0–100)^d [Punkte]	Tatsächliche Änderung bei Response ≥ 10 Punkte^e [Punkte]	Tatsächliche Änderung bei Response ≥ 15 Punkte^f [Punkte]
Dyspnoe, Schlaflosigkeit, Appetitverlust, Verstopfung, Diarrhö (jeweils 1 Item, Wertebereich 1–4)	3	3	33,3	33,3	33,3
Rollenfunktion, kognitive Funktion, soziale Funktion, Schmerzen, Übelkeit und Erbrechen (jeweils 2 Items, Wertebereich 1–4)	3	6	16,7	16,7	16,7
Globaler Gesundheitszustand (2 Items, Wertebereich 1–7)	6	12	8,3	16,7	16,7
Fatigue (3 Items, Wertebereich 1–4)	3	9	11,1	11,1	22,2
Emotionale Funktion (4 Items, Wertebereich 1–4)	3	12	8,3	16,7	16,7
Körperliche Funktion (5 Items, Wertebereich 1–4)	3	15	6,7	13,3	20,0
a. Die Skala „finanzielle Schwierigkeiten“ (1 Item) wird in der Nutzenbewertung nicht betrachtet. b. Die Items können jeweils Werte von 1 bis 4 annehmen (bzw. von 1 bis 7 bei den Items der Skala zum globalen Gesundheitszustand). Der rohe Wert der Skala ergibt aus der Summe der Einzelitems geteilt durch die Anzahl der Einzelitems. c. Die Anzahl möglicher Änderungsschritte der Skala ergibt sich aus der Anzahl der möglichen Änderungsschritte je Item multipliziert mit der Anzahl der Items. d. Der kleinstmögliche Änderungsschritt auf der normierten Skala ergibt sich aus der Division der Skalenspannweite der normierten Skala (0–100) durch die Anzahl der möglichen Änderungsschritte der Skala. e. tatsächliche Änderung auf der Skala bei Anwendung einer Responseschwelle von mindestens 10 Punkten f. tatsächliche Änderung auf der Skala bei Anwendung einer Responseschwelle von mindestens 15 Punkten					

Daraus ergibt sich, dass sich lediglich für 2 der insgesamt 14 Skalen des EORTC QLQ-C30 bei Anwendung der beiden Responseschwellen 10 bzw. 15 Punkte in einer Responderanalyse unterschiedliche Anzahlen an Respondern ergeben können, während sie für die übrigen

Skalen zu identischen Responderzahlen führen. So bedeutet beispielsweise für die 5 Skalen, die lediglich 1 Item enthalten, die jeweils kleinste mögliche Änderung der Skala eine Änderung um 33,3 Punkte auf der von 0 bis 100 normierten Skala. Dies führt dazu, dass diese Patientinnen und Patienten sowohl in einer Auswertung mit einer Responseschwelle von 10 Punkten als auch einer Auswertung mit einer Responseschwelle von 15 Punkten gleichermaßen als Responder in die Analyse eingehen. Die Anzahl der Responder ist damit in beiden Auswertungen identisch. Dies betrifft gleichermaßen die 6 Skalen mit 2 Items (kleinstmöglicher Änderungsschritt 16,7 Punkte) wie auch die Skala mit 4 Items (Änderungsschritt 16,7 Punkte, der kleinstmögliche Schritt von 8,3 Punkten erfüllt keine der beiden Responseschwellen). Lediglich bei den Skalen zu Fatigue (3 Items) und der körperlichen Funktion (5 Items) ergeben sich Unterschiede: Für die 3-Item-Skala entspricht eine Responseschwelle von 10 Punkten einem Änderungsschritt auf einem Item der Skala (11,1 Punkte) während eine Responseschwelle von 15 Punkten 2 Änderungsschritte erfordert (22,2 Punkte). Ähnlich liegen die entsprechenden Werte für die 5-Item-Skala bei 2 Änderungsschritten für die 10-Punkte-Schwelle (13,3 Punkte) und bei 3 Änderungsschritten für die 15-Punkte-Schwelle (20,0 Punkte). Insgesamt ergeben sich daher für 12 (86 %) von 14 Aussagen, die basierend auf Responderanalysen der Skalen des EORTC QLQ-C30 getroffen werden, mit beiden Responseschwellen identische Ergebnisse. Unterschiede können sich lediglich für die 2 genannten Skalen ergeben. Allerdings besteht der Unterschied zwischen den beiden Responseschwellen in nur 1 Schritt auf der jeweiligen Skala. In dieser spezifischen Situation des EORTC wird daher die Auswertung mit einer Responseschwelle von 10 Punkten als hinreichende Annäherung an eine Auswertung mit einer 15 %-Schwelle (15 Punkte) betrachtet und für die Nutzenbewertung herangezogen.

Die oben gemachte Betrachtung lässt sich auch auf die indikationsspezifischen Zusatzmodule des EORTC übertragen, die ausschließlich in Kombination mit dem EORTC QLQ-C30 zum Einsatz kommen. So umfasst das Zusatzmodul zum hepatozellulären Karzinom (HCC18) insgesamt 8 Skalen, von denen 6 Skalen jeweils 1 bzw. 2 Items enthalten und damit zu beiden Responseschwellen identische Ergebnisse liefern. Lediglich 2 Skalen (Fatigue, 3 Items sowie Ernährung, 5 Items) unterscheiden sich – wie oben dargestellt – bei Anwendung der beiden Responseschwellen um einen Änderungsschritt auf der Skala.

Insgesamt sind damit mehr als 70 % der getroffenen Aussagen identisch und bei den abweichenden Skalen bedeutet der Unterschied in keinem Fall mehr als 1 Änderungsschritt auf der Skala. Damit gilt auch für die gemeinsame Anwendung des EORTC QLQ-C30 mit dem HCC18, dass die Auswertung mit einer Responseschwelle von 10 Punkten als hinreichende Annäherung an eine Auswertung mit einer 15 %-Schwelle betrachtet und für die Nutzenbewertung herangezogen wird.

Diese Betrachtung lässt sich auch auf die Mehrzahl der indikationsspezifischen Zusatzmodule des EORTC übertragen, die ausschließlich in Kombination mit dem EORTC QLQ-C30 zum Einsatz kommen (siehe auch [11]).

C.2 Herleitung der Responsekriterien für den Physical Component Summary (PCS) und Mental Component Summary (MCS) des SF-36

Bei den Summenscores des SF-36 (Physical Component Summary [PCS] und Mental Component Summary [MCS]) handelt es sich nach Transformation um normierte Werte, die nicht mehr eine Skalenspannweite von 0 bis 100 haben. Wie sich in dieser besonderen Situation von normierten Werten die Skalenspannweite bestimmen lässt, wird nachfolgend erläutert. Dazu wird zunächst die Bildung des Physical Component Summary (PCS) und des Mental Component Summary (MCS) skizziert (siehe Kapitel 5 des Manuals zum SF-36 [38]) und dann eine Skalenspannweite als Grundlage für die Bestimmung eines Responsekriteriums angegeben. Alle Angaben gelten für Version 2 des SF-36 mit einer Recall-Zeit von 4 Wochen.

Der SF-36 besteht aus 36 Items, von denen 35 Items den in Tabelle 11 aufgelisteten 8 Domänen zugeordnet sind. Die Bildung der Domänen-Scores erfolgt durch Summierung der Werte der Items pro Domäne und anschließender Transformation der Summen auf Werte zwischen ≥ 0 und ≤ 100 . Anschließend werden die Domänen-Scores für jede Teilnehmerin / jeden Teilnehmer auf Basis der Mittelwerte und Standardabweichungen der Normstichprobe 2009 des SF-36 (siehe Kapitel 14 im Manual [38]) in einen z-Score transformiert. Die z-Scores werden anschließend durch Multiplikation mit dem Faktor 10 und Addition von 50 in T-Scores umgewandelt. Diese Normierung führt dazu, dass sich für eine Normalbevölkerung (repräsentiert durch die Normstichprobe des SF-36) eine Verteilung mit Mittelwert 50 und Standardabweichung 10 ergibt. Ein T-Score unter 50 bedeutet eine schlechtere Funktion oder einen schlechteren Zustand verglichen mit dem jeweiligen mittleren Wert der Normalbevölkerung; bei einem T-Score über 50 ist es umgekehrt. Diese Berechnungen führen dazu, dass die Skalenspannweite der 8 Domänen nicht mehr 100 beträgt.

Zur Berechnung der beiden übergeordneten Component Summaries (PCS und MCS) werden die z-Scores der 8 Domänen jeweils mit den in Tabelle 11 genannten Koeffizienten gewichtet und addiert. Die z-Scores der beiden Component Summaries werden wie die z-Scores der Domänen anschließend durch Multiplikation mit dem Faktor 10 und Addition von 50 in T-Scores umgewandelt. Wie für die 8 Domänenscores führt diese Normierung dazu, dass sich für eine Normalbevölkerung (repräsentiert durch die Normstichprobe des SF-36) eine Verteilung mit Mittelwert 50 und Standardabweichung 10 ergibt. Ein T-Score unter 50 bedeutet eine schlechtere gesundheitsbezogene Lebensqualität verglichen mit der mittleren gesundheitsbezogenen Lebensqualität der Normalbevölkerung; bei einem T-Score über 50 ist es umgekehrt.

Tabelle 12: Anzahl Items pro Domäne und Gewichtung der Domänen in den Component Summaries PCS und MCS

Domäne (Anzahl Items) ^a	Koeffizienten	
	PCS	MCS
körperliche Funktionsfähigkeit (10)	0,42402	-0,22999
körperliche Rollenfunktion (4)	0,35119	-0,12329
körperlicher Schmerz (2)	0,31754	-0,09731
allgemeine Gesundheitswahrnehmung (5)	0,24954	-0,01571
Vitalität (4)	0,02877	0,23534
soziale Funktionsfähigkeit (2)	-0,00753	0,26876
emotionale Rollenfunktion (3)	-0,19206	0,43407
psychisches Wohlbefinden (5)	-0,22069	0,48581
a. nur 35 der 36 Items werden zur Berechnung des PCS und des MCS verwendet.		
MCS: Mental Component Summary; PCS: Physical Component Summary		

Wie aus Tabelle 12 hervorgeht, sind die Koeffizienten je Domäne und Component Summary (PCS, MCS) positiv oder negativ. Der Einbezug von negativen Koeffizienten führt dazu, dass die Minima und Maxima des PCS und MCS nicht berechnet werden können, indem allen Items jeweils der beste oder der schlechteste Wert zugeordnet wird (Szenario 1 in Tabelle 13). Wird die Skalenspannweite dagegen berechnet, indem die Items jeweils so angekreuzt werden, dass der PCS bzw. MCS maximiert bzw. minimiert werden, ergibt sich eine größere Skalenspannweite (Tabelle 13, Szenario 2 berechnet mittels Normstichprobe 1998 und Szenario 3 berechnet mittels Normstichprobe 2009 [Personal communication mit Jakob Bue Bjørner, MD, PhD, Chief Science Officer at QualityMetric [39]] [40]). Neben den berechneten möglichen Skalenspannweiten (Szenarien 1, 2 und 3) liegen im Manual des SF-36 [38] die empirischen Minima und Maxima in der Normstichprobe von 2009 vor (Tabelle 13, Szenario 4).

Tabelle 13: Minima und Maxima des PCS und MCS des SF-36 in verschiedenen Szenarien

	Szenario 1: PCS- und MCS-Ergebnisse unter Annahme jeweils niedrigster bzw. höchster möglicher Werte pro Item (berechnet mittels Normstichprobe 1998)^a		Szenario 2: Minimierung bzw. Maximierung des PCS bzw. MCS (berechnet mittels Normstichprobe 1998)^a		Szenario 3: Minimierung bzw. Maximierung des PCS bzw. MCS (berechnet mittels Normstichprobe 2009)^b		Szenario 4: Beobachtete Werte in der Normstichprobe 2009^c	
	Min	Max	Min	Max	Min	Max	Min	Max
PCS	22,4	59	0,7	80,7	5,0	79,8	7,3	70,1
MCS	10,6	62,3	-8,8	81,6	-3,3	80,1	5,8	69,9

a. eigene Berechnung
b. Personal communication mit Jakob Bue Bjørner, MD, PhD, Chief Science Officer at QualityMetric [39,40]
c. Angabe in Table 7.1 des Manuals des SF-36 [38]
Max: Maximum; MCS: Mental Component Summary; Min: Minimum; PCS: Physical Component Summary

Von den in Tabelle 13 bestimmten theoretischen Minima und Maxima (Szenarien 1, 2 und 3) sind die Werte gemäß Szenario 1 als Grundlage zur Bestimmung eines Responsekriteriums nicht geeignet, da sie nicht die maximal mögliche Skalenspannweite umfassen. Die Minima und Maxima gemäß Szenarien 2 und 3 umfassen zwar die maximal mögliche Skalenspannweite; die Minima und Maxima treten aber wegen der Gewichtung der Komponenten der Summenscores nur auf, wenn eine maximale mentale Beeinträchtigung mit einer minimalen körperlichen Beeinträchtigung (und umgekehrt) einhergeht.

Wegen dieses Auswertungsalgorithmus wird für die beiden Component Summaries des SF-36 die Skalenspannweite der empirischen Minima und Maxima aus der Normstichprobe von 2009 (Szenario 4) für die Bestimmung des Responsekriteriums (15 % der Skalenspannweite) verwendet. Bei einem Minimum von ca. 7 (PCS) bzw. 6 (MCS) und einem Maximum von ca. 70 für beide Domänen (siehe Tabelle 13) resultiert eine Skalenspannweite von ca. 63 (PCS) bzw. 64 (MCS), wodurch sich bei 15 % der Skalenspannweite ein Responsekriterium von knapp 10 Punkten für die beiden Component Summaries ergibt (PCS: 9,4; MCS: 9,6). Somit stellen 10 Punkte für den PCS und MCS des SF-36 jeweils eine geeignete Annäherung für ein Responsekriterium unter Verwendung von 15 % der Skalenspannweite dar.

Für den SF-36 mit einer Recall-Zeit von 1 Woche liegen laut Manual (Table 7.1) [38] die folgenden Minima und Maxima vor: PCS: 10,8 und 75,5; MCS 5,6 und 69,7. Daraus ergeben sich die Skalenweiten von ca. 64,7 und 64,1 und die Responsekriterien von 9,7 und 9,6. Auch hier stellen 10 Punkte für den PCS und MCS des SF-36 jeweils eine geeignete Annäherung für ein Responsekriterium unter Verwendung von 15 % der Skalenspannweite dar (siehe auch [12]).

C.3 PROMIS / Item-Banken

PROMIS-Itembanken werden über ein computer-adaptives Testen (CAT) angewendet [41]. Bei einem CAT sind die Anzahl und der Schweregrad der Items nicht vordefiniert, sondern ergeben sich aus den jeweiligen Antworten der Patientinnen und Patienten. Zudem werden die Rohwerte in T-Scores transformiert, die (zumindest theoretisch) von minus unendlich bis unendlich reichen.

Für Daten, die mit dem PROMIS-System erhoben wurden, wurde ein Vorschlag für ein Responsekriterium mit David Cella, Devin Peipert und Kollegen diskutiert [42]. Auf der Grundlage der Skalenverteilungsdaten von HealthMeasures kann ein funktionaler Skalenbereich von 20 bis 80 angenommen werden, da nur sehr wenige Patientinnen und Patienten T-Scores unter 20 oder über 80 aufweisen. Auf der Grundlage dieses Skalenbereichs entspricht ein Schwellenwert von 9 T-Score-Punkten 15 % der Skalenspannweite.