Responder analysis thresholds for complex scales in benefit assessments – the 15% approach[1]

WORKING PAPER

Project: GA25-04 Version: 1.0 Status: 13 Jul 2026 DOI: 10.60584/GA25-04_en

[1] Translation of the working paper *Schwellenwerte für Responderanalysen bei komplexen Skalen in der Nutzenbewertung – der 15 %-Ansatz* (Version 1.0; Status: 13 July 2026). Please note: This translation is provided as a service by IQWiG to English-language readers. However, solely the German original text is absolutely authoritative and legally binding.

Publishing details

Publisher

Institute for Quality and Efficiency in Health Care

Topic

Responder analysis thresholds for complex scales in benefit assessments – the 15% approach

Commissioning agency

Prepared in the context of IQWiG's general commission

Internal Project No.

GA25-04

DOI-URL

https://doi.org/10.60584/GA25-04_en

Address of publisher

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

Phone:+49 221 35685-0
Fax: +49 221 35685-1
Email: info@iqwig.de
Internet: www.iqwig.de

Recommended citation

Institute for Quality and Efficiency in Health Care. Responder analysis thresholds for complex scales in benefit assessments – the 15% approach; Working paper [online]. 2026 [Accessed: DD.MM.YYYY]. URL: https://doi.org/10.60584/GA25-04_en.

Keywords

Patient Reported Outcomes

This working paper was prepared without collaboration with external experts.

The responsibility for the contents of the report lies solely with IQWiG.

IQWiG employees

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

Executive summary

In 2020, IQWiG proposed a new approach to determining responder analysis thresholds for complex scales, particularly those used to assess patient-reported outcomes (PROs), and published it for consultation. Since the publication of the General Methods Version 6.0, this approach, known as the 15% approach, has been applied.

The rationale for addressing response thresholds was that, in recent years, responder analyses – which employ a response criterion in the form of an individual minimal important difference (MID) – have increasingly been used to evaluate medical interventions. However, determining an MID is associated with methodological problems, and for individual instruments, a multitude of MIDs are often published. This means that no uniform MIDs can be identified for scales; rather, they depend on the specific patient group under consideration, their disease, the severity of that disease, as well as the methods used to determine the MID. Although there are now approaches available to evaluate the quality of validation studies on MIDs, it is evident that publications of studies for determining an MID often fail to report relevant aspects of the methods or criteria for an evaluation.

The aim was therefore to identify a generic response threshold that represents a change perceptible to patients with sufficient certainty.

To this end, systematic reviews on MIDs were identified through a focused search, and the MIDs reported therein were compiled. In total, the MIDs reported in 8 of the 16 identified systematic reviews were extracted. These cover a broad range of topics, such as orthopaedic and oncological indications, paediatrics, or instruments for measuring health-related quality of life, as well as specific concepts (fatigue) and diseases (chronic obstructive pulmonary disease, rheumatism). In a subsequent step, the identified MIDs were related to the scale range of the respective measurement instrument.

As a result, a value of 15% of the scale range of the respective scale was determined as the response criterion. This response criterion is considered a plausible value for a relatively small, but sufficiently certain, individual change on a scale that falls within the perceptible range. This is explicitly not a response criterion in terms of an individual MID. Even though the 15% threshold is intended to represent a relatively small change, in many cases it may lie above identified MIDs.

IQWiG first introduced the approach in 2020 as part of its General Methods Version 6.0 and discussed it with the scientific community during the commenting procedure and in the oral hearing. In 2021, the Federal Joint Committee (G-BA) incorporated the new threshold into the dossier templates for assessing the added benefit of new drugs. The G-BA also conducted a

commenting procedure and an oral hearing on this matter. Since then, the application of the 15% threshold has become an established approach in the national process for the benefit assessment of new drugs. The aim of this working paper is to bring together in a single document the methods for deriving the generic response threshold, the discussion that took place during the commenting procedure for the General Methods Version 6.0, and further specifications for specific instruments. This summary is intended to make the available information accessible to a broader readership.

Table of contents

List of tables

Page

List of figures

Page

List of abbreviations

Abbreviation	Meaning
CAT	computer-adaptive testing
COPD	chronic obstructive pulmonary disease
EORTC	European Organization for Research and Treatment of Cancer
G-BA	Gemeinsamer Bundesausschuss (Federal Joint Committee)
GRC	Global Rating of Change
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (Institute for Quality and Efficiency in Health Care)
MCS	Mental Component Summary
MID	minimal important difference
PCS	Physical Component Summary
QLQ-C30	Quality of Life Questionnaire – Core 30
PRO	patient-reported outcome
RCT	randomized controlled trial
ROC	receiver operating characteristic
SF-36	Short Form 36
SGRQ	St. George's Respiratory Questionnaire
WOMAC	Western Ontario and McMaster Osteoarthritis Index

1 Background

In 2020, IQWiG proposed a new approach to determining responder analysis thresholds for complex scales, particularly those used to assess patient-reported outcomes (PROs), and published it for consultation. Since the publication of the General Methods Version 6.0 [1], this approach, known as the 15% approach, has been applied.

The rationale for addressing response thresholds was that, in recent years, responder analyses – which employ a response criterion in the form of an individual minimal important difference (MID) – have increasingly been used to evaluate medical interventions. An MID is defined as the smallest difference in an instrument's score that is perceptible to a patient and is evaluated as an improvement or deterioration [2,3]. However, determining an MID is associated with methodological problems, and for individual instruments, a multitude of MIDs are often published [4]. This means that no uniform MIDs can be identified for scales; rather, they depend on the specific patient group under consideration, their disease, the severity of that disease, as well as the methods used to determine the MID. Although there are now approaches available to evaluate the quality of validation studies on MIDs, it is evident that publications of studies for determining an MID often fail to report relevant aspects of the methods or criteria for evaluation [5].

The aim of the analysis of 2020 was therefore to identify a generic response threshold that, while on the one hand is rather small, is at the same time within the range perceptible to the majority of patients with sufficient certainty. A generic threshold of 10% of the instrument's scale range was discussed as a starting point, as this has been discussed in the scientific literature on various occasions as a "rough" and context-independent approximation of an MID [6,7]. To develop the new approach for determining a generic response criterion, IQWiG used a focused search to identify systematic reviews on MIDs, compiled the MIDs reported therein, and set them in relation to the scale range of the respective instrument.

2 Research question

The aim of the analysis of 2020 was to identify a generic response threshold for responder analyses for complex scales – particularly those designed to measure PROs – that represents a change perceptible to patients with sufficient certainty. While this threshold is intended to represent a relatively small change, it is explicitly not intended to identify a “generic” MID. Therefore, it may lie above previously identified MIDs for a given instrument.

3 Course of the project

In 2020, with the General Methods Version 6.0, IQWiG published a new approach to determining responder analysis thresholds for complex scales, particularly those used to measure PROs [1], and discussed it with experts during the commenting procedure and the oral hearing [8]. Since then, the response threshold specified in the General Methods of 15% of the scale range of the respective instrument has been applied. The Federal Joint Committee (G-BA) subsequently incorporated the specifications regarding response thresholds into its dossier templates and conducted a commenting procedure on this matter [9,10].

The methodological basis for determining a threshold has so far been published exclusively as part of the documentation and evaluation of the hearing on the aforementioned methods paper. Further specifications for specific instruments are also described in individual IQWiG dossier assessments for the assessment of the added benefit of new drugs [11,12].

This working paper brings together the aspects published in various places into a single document and additionally includes a more detailed presentation of the methods and results as the basis for determining a threshold in 2020. This working paper documents the procedure followed at that time and does not include an update of the underlying literature.

4 Methods

The methodological approach for the 15% threshold published in 2020 with the General Methods Version 6.0 [1] is described below.

Information retrieval

A focused search for systematic reviews for MIDs on PROs was conducted in a bibliographic database (MEDLINE) (search date: 22 May 2019). The reviews were selected based on the following inclusion criteria:

- Reviews published in 2016 or later.
- Reviews including studies in which MIDs were empirically derived.
- Reviews including studies in which MIDs were determined for instruments that are patient-reported, comprise more than one item, and whose scores are not directly interpretable (abstract scales as opposed to, for example, natural scales whose results are directly interpretable).

Selection of systematic reviews

The included systematic reviews were to provide as broad an overview as possible of MIDs for PROs across various therapeutic indications and populations. Systematic reviews were not included in the analysis if the studies largely overlapped with those in the already included reviews.

Extraction of MIDs

The full texts of the citations included in the systematic reviews were obtained. The publications were reviewed with regard to the methods used to determine the respective MID. The methods for determining the MID had to meet the following criteria:

- longitudinal study
- anchor-based MID
- patient-reported anchor (or, for children/patients with dementia, reported by parents or a proxy)
- “Global Rating of Change” (GRC) anchor
- The cut-off for the GRC anchor was to be minimal, small, little, or at most moderate to ensure the determination of an MID.
- The scales had to be abstract PRO multi-item scales.

Data on the empirically derived MIDs were extracted only if the methods described in the publications met the above criteria. In addition, the respective scale range of the instrument was identified, and the MID was put in relation to it.

5 Results

5.1 Results of information retrieval

Figure 1 shows the results of the focused information retrieval and study selection according to the criteria for study inclusion. The search strategy for searching bibliographic databases can be found in Appendix B. The search was conducted on 21 May 2019.

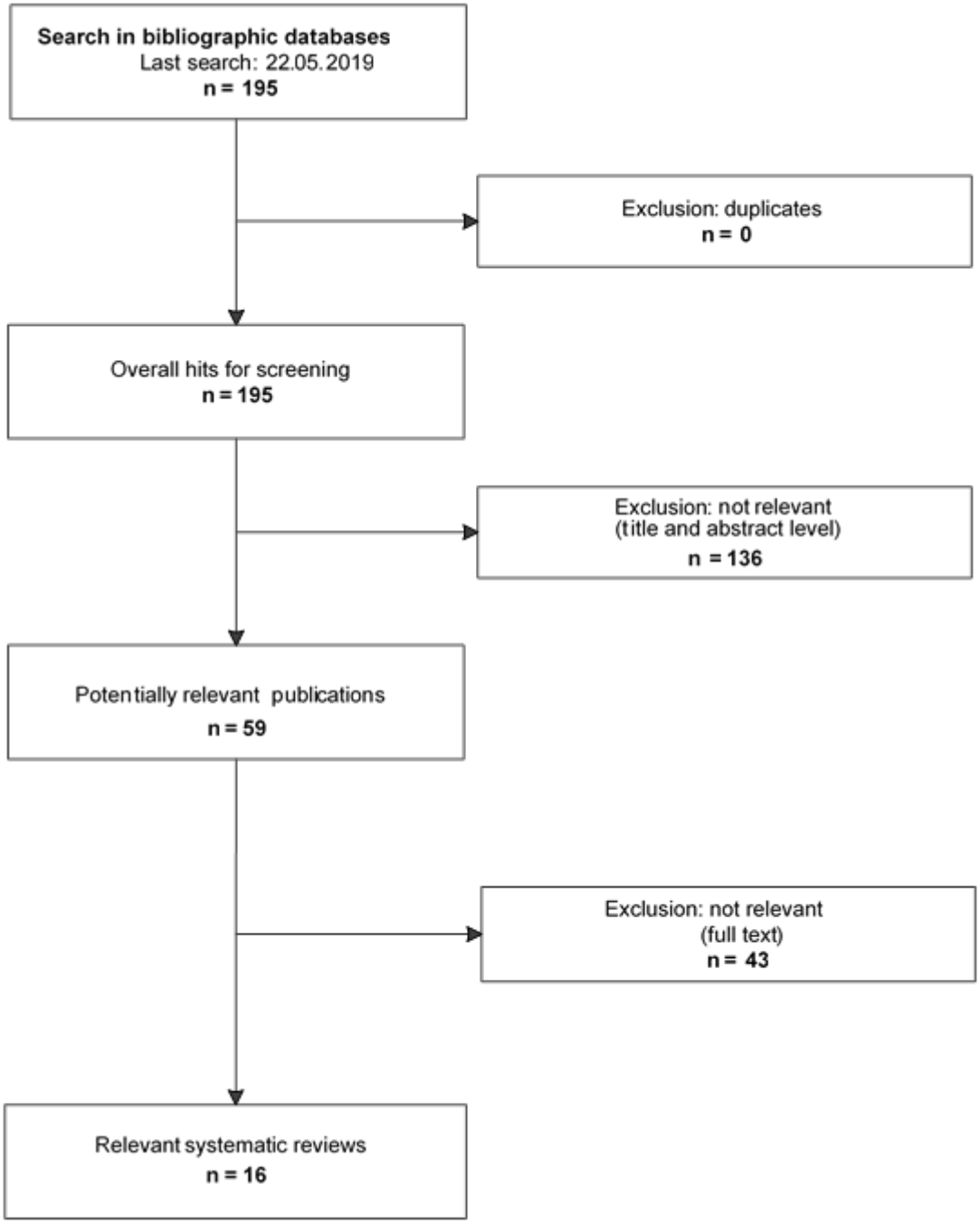

Figure 1: Results of the focused information retrieval for systematic reviews on MIDs

The literature search identified 16 review articles [13-28], of which 8 [13,19,20,22,23,25,27,28] were included in the analysis. These included studies on orthopaedic [22,28], paediatric [20], and oncological [27] indications, as well as studies on specific conditions such as rheumatism [19] or chronic obstructive pulmonary disease (COPD) [13]. One of the reviews focused exclusively on studies for determining the MID of instruments for assessing fatigue [25], while another focused on questionnaires for assessing health-related quality of life [23].

We did not extract data from the remaining 8 reviews [14-18,21,24,26], as these exclusively covered orthopaedic indications and the studies largely overlapped with those in the reviews already included. Thus, no information beyond that already extracted from the other reviews was expected.

5.2 Results on MIDs from the studies included in the systematic reviews

The extracted reviews are briefly characterized below.

- Doganay 2016 [19] (rheumatology): The aim was to evaluate MIDs in rheumatology over the past 2 decades. Taking the inclusion criteria into account, MIDs were extracted for 17 instruments from a total of 15 studies. The range of MIDs extends from < 1% to 38% of the scale range of the respective instrument. An additional 21 of the studies identified by the review did not meet the inclusion criteria.
- St-Pierre 2016 [28] and Hao 2019 [22] (orthopaedics): The St-Pierre review examined the psychometric properties of PROs for evaluating symptoms and functional limitations in patients with rotator cuff injuries. Data from 8 of 11 studies on 8 PRO instruments met the inclusion criteria for the present analysis. The range of MIDs extends from 6% to 23%. The aim of the Hao 2019 review was to determine anchor-based MIDs for PROs in patients with shoulder complaints. A total of 11 out of 22 studies met the inclusion criteria, and MIDs were extracted for 9 different measurement instruments. The range of MIDs across all instruments is 4% to 23% of the scale range.
- Ebrahim 2017 [20] (paediatrics): The review examined anchor-based MIDs for PROs in paediatric studies of various conditions. Data on MIDs were extracted from 7 included studies for 10 measurement instruments. The range of MIDs extends from 2% to 34% of the scale range. An additional 25 of the studies identified by the review did not meet the inclusion criteria.
- Nordin 2016 [25] (fatigue): The aim was to compile and evaluate published MIDs for PROs related to fatigue. MID data from 4 included studies were considered for 5 PROs. The range of MIDs extends from 2% to 24% of the scale range. An additional 37 of the studies identified by the review did not meet the inclusion criteria.

- Jayadevappa 2017 [23] (health-related quality of life): The aim was to evaluate MIDs in the assessment of health-related quality of life using generic and disease-specific instruments. To this end, instruments across various therapeutic indications were examined. MIDs were extracted from 23 studies for 24 instruments. The range of MIDs extends from < 1% to 38% of the scale range. An additional 57 of the studies identified by the review did not meet the inclusion criteria.
- Alma 2018 [13] (COPD): For this review, the available evidence on MIDs of PRO instruments in the indication of COPD was first identified. The aim was to determine the average MID for each PRO instrument based on anchor- and distribution-based MIDs using method triangulation. For the present analysis, the MIDs for 4 instruments were extracted from a total of 5 studies. The range extends from 3% to 17% of the scale range. An additional 16 of the studies identified by the review did not meet the inclusion criteria.
- Ousmen 2018 [27] (oncology): The aim was to examine the statistical methods used to determine the MID for PROs in oncology. For 2 instruments, the MIDs were derived from 2 studies. The range extends from < 1% to 27% of the scale range. An additional 44 of the studies identified by the review did not meet the inclusion criteria.

The systematic reviews included between 21 [13] and 80 [23] studies on MIDs. The majority of these studies did not meet the inclusion criteria for the present analysis. The most common reason for exclusion was that a GRC anchor was not used to derive the MID. Other reasons for exclusion included: the methods for determining the MID was described unclearly; the anchor for the MID was not patient-reported; the cut-off for the GRC anchor was greater than minimal; the MID was determined using a distribution-based method; or the studies were cross-sectional.

The following Table 1 shows the distribution of the extracted MIDs in relation to the scale range in %. A more detailed overview based on the individual reviews and instruments can be found in Appendix B.

Table 1: Range of empirically derived MIDs relative to the range of the respective scale used, in %

Review	Therapeutic indication	Range of the reported MIDs relative to the range of the scale in %
Doganay 2016 [19]	Rheumatology	< 1–38
St-Pierre 2016 [28]	Orthopaedics	6–23
Hao 2019 [22]	Orthopaedics	4–23
Ebrahim 2017 [20]	Paediatrics	2–34
Nordin 2016 [25]	Fatigue	2–24
Jayadevappa 2017 [23]	▪ Oncology ▪ Cardiovascular diseases ▪ Respiratory diseases ▪ Diseases of the urogenital tract ▪ Musculoskeletal diseases ▪ Paediatrics ▪ Miscellaneous	3–18 9–16 3–17 6–18 1–23 < 1–38 < 1–23
Alma 2018 [13]	COPD	3–17
Ousmen 2018 [27]	Oncology	< 1–27
COPD: chronic obstructive pulmonary disease; MID: minimal important difference		

Across all therapeutic indication, an analysis of empirically derived MIDs reveals a considerable variability in the results of studies for determining MIDs.

The MID values range from < 1% to 38% of the scale range (see Table 3 to Table 10 in Appendix B). It should be emphasized that this list already includes only those studies that met the specified minimum quality criteria (e.g., patient-reported anchor, longitudinal study, see above). These criteria constitute an essential component of the instrument proposed by the McMaster University group for assessing the quality of MIDs [5].

Despite meeting the established minimum quality criteria, the methods for determining MIDs varies vastly across studies. In the included studies, MIDs were predominantly determined either as the mean change in the score of those patients who had experienced a (minimal) noticeable change compared to baseline or using receiver operating characteristic (ROC) curves. In some cases, MIDs were determined separately for improvement and deterioration, for different patient groups, for different recall times, as well as for different domains or the total score of an instrument.

The variability of the MIDs derived from the various approaches is described below using the Western Ontario and McMaster Osteoarthritis Index (WOMAC) as an example. The WOMAC consists of 24 items that are assigned to the domains of pain, stiffness, and physical function. Scores are calculated for the individual domains and a total score. In the Doganay review of

2016 [19], MIDs for the WOMAC were reported in 11 studies. Of these, 8 studies met the inclusion criteria for the present analysis. This resulted in a wide range of different MIDs. These are presented in the following Table 2.

Table 2: Summary of MIDs for the WOMAC from 8 studies included in the Doganay review [19]

Instrument	Domain	MIDs as a % of the scale range	% range of MIDs
Western Ontario and McMaster Osteoarthritis Index (WOMAC)[a]	Pain	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
	Stiffness	2.9; 5.1; 6.3; 7.2; 10.1; 14.5; 18.8; 25.9; 33.2	2.9–33.2
	Physical function	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
	Total 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: Scale range of the instrument: 0–10 points. MID: minimal important difference; WOMAC: Western Ontario and McMaster Osteoarthritis Index			

In 5 of these studies, the MIDs for the WOMAC were determined as the mean change compared to baseline, and in 2 studies using ROC curves. In another study, the MIDs were determined and reported using both methods. In 2 studies, MIDs for improvement and deterioration were calculated for each domain and for the total score. In the studies, the MIDs were determined at different recall times (ranging from 2 months to 24 months after baseline). In some cases, MIDs were determined for both the individual domains and the total score. For 1 study, only the MID for the total score was reported; in 2 other studies, only the MIDs for individual domains were reported, but not for the total score.

These results once again clearly illustrate the situation described at the outset, namely that no uniform MIDs can be identified for individual instruments. Although minimum methodological standards were already established in the present analysis through the inclusion criteria, the values differ considerably. Added to this are differences due to the patient group, the disease, the severity of the disease, or the direction of the change.

When viewed in the context of all results, the values at the lower end of the range of MIDs reported across all studies appear unrealistically small and are not plausible in terms of distinguishing between a perceptible change and measurement uncertainty. The question of whether the values at the upper end of the spectrum are more reliable cannot be answered due to the lack of accepted standards for quality assessment, as well as the poor reporting quality of MID studies (see below). However, a systematic examination of the extracted MIDs shows that the majority of empirically derived MIDs lie above the generic threshold of 10% discussed in the literature, but mostly below 20% of the respective scale range. Based on the

results of this systematic examination of MIDs, a threshold of 15% was therefore considered plausibly suitable for reliably representing a change perceptible to patients. This response threshold is intended to represent a relatively small change, but the value may well lie above a minimal threshold. This means that determining this threshold is explicitly not about identifying a response criterion in terms of an individual MID. The 15% threshold was set pragmatically based on the MIDs identified in the systematic reviews, rather than as a value calculated from the results. The methods for identifying MIDs, associated quality criteria, and related approaches to analysing PRO instruments are currently the subject of scientific debate. In this context, the threshold determined is intended to serve as a pragmatic solution that provides clarity for all stakeholders, particularly in the context of early benefit assessment.

6 Discussion

Setting a 15% threshold is a pragmatic decision made in light of the lack of uniform instrument-specific response thresholds in terms of an individual MID, as well as the methodological shortcomings of previously identified MIDs described above. Furthermore, publications on MIDs often lack sufficient information to assess their quality [4], making it impossible to currently identify the most reliable value for a single instrument from the multitude of available MIDs.

The IQWiG methods paper further stipulated that analyses with a response threshold of 15% of the scale range are accepted even if they were not specified in the study protocol and analysis plan (post-hoc analysis of the respective study). Since this analysis is required (predefined) for IQWiG's benefit assessment, the risk of selective reporting – which otherwise exists with post-hoc response criteria – is reduced. This step has created reliability for all stakeholders, particularly regarding the process of the early benefit assessment of drugs, with regard to which analyses are accepted.

6.1 Commenting procedure on the proposed approach

The new threshold was first published with the draft of the General Methods Version 6.0 at the end of 2019, discussed with experts during the subsequent commenting procedure, and established in 2020 with the final Version 6.0 [1,8]. In 2021, the G-BA incorporated the new threshold into the dossier templates for the assessment of the added benefit of new drugs. The G-BA also conducted a commenting procedure on this matter, including an oral hearing [9,10,29]. The following aspects were discussed in particular during the proceedings.

Generic response threshold versus MID

During the hearing, there was discussion as to whether the response criterion of 15% of the scale range was too vague and too high to constitute an MID. However, it should be noted that – as previously described – the new threshold is not intended to serve as an individual MID. Rather, the threshold is intended to reliably reflect a change that is perceptible to patients (see above). It is assumed that the value may well lie above the respective MID. This characteristic also accounts for the existing variability of MIDs and helps ensure that the response criterion does not fall below the MID too frequently in different scenarios. At the same time, the value is small enough to reflect a relatively minor change.

In this context, it is correct that the definition of the response threshold as 15% of the scale range means that some of the response criteria used previously are no longer considered. This applied, for example, to the response criteria cited by the commenters for the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire – Core 30 (QLQ-C30) of 10 points or the St. George's Respiratory Questionnaire (SGRQ) of 4 points (scale range of the instrument 0 to 100 points).

To assess this issue, however, it is necessary to examine the quality of the response criteria in question. In the examples cited by the commenters, the response criteria generally stem from studies that do not reflect the current state of scientific knowledge regarding studies for determining MIDs.

For the 4-point MID for the SGRQ, a narrative review from 2005 is primarily cited [30]; other publications from 2015 and 2018 describe MIDs ranging from 6 to 9 points [13,31].

The response criterion for the EORTC QLQ-C30 is primarily justified by the work of Osoba et al. [32]. In the dataset used there, the correlation between the anchor and the scale, for example, did not meet the requirements currently under discussion. In this situation, the EORTC, which is responsible for the instrument, initiated a project to determine MIDs for the EORTC QLQ-C30 questionnaire [33]. Initial results regarding MIDs for group differences are now available; results regarding MIDs at the individual patient level are still pending. In the available publication, the authors discuss that no uniform MID can be determined for group differences, but rather that these vary between the individual scales as well as across different disease settings [34] (for the current approach to individual thresholds for the EORTC QLQ-C30 and its supplementary modules, see Appendix C.1).

Robust effect estimation

Furthermore, the pharmaceutical companies that submitted comments in particular expressed concern that the proposed approach would raise the bar and make it more difficult to demonstrate the added benefit of interventions. The commenters did not explain why they assume that the proposed response criterion would make it more difficult to demonstrate added benefit, nor did they submit data supporting this assumption during the commenting procedure of either IQWiG or the G-BA. In general, it cannot be assumed that (robust) effects are no longer detectable if alternative response criteria fall within the lower range of a scale. This assessment was confirmed by the discussion in the scientific hearing on IQWiG‘s methods paper.

Substantive evaluation for natural scales

Another point of discussion concerned the applicability of the threshold to natural scales that have no range, such as the 6-minute walk test. It is true that the proposed response criterion of at least 15% of the scale range cannot be applied, for example, to open-ended scales. However, such an application is not planned. Determining a threshold addresses complex scales (particularly psychometric scales) whose results cannot be readily interpreted. The proposed response criterion of 15% of the scale range is not intended for use with scales whose results are expressed in natural units and which are therefore amenable to a content-based assessment of the relevance of an effect. Since complex psychometric scales always

have a maximum value, the problem cited by the commenters cannot arise (for handling special cases, see, for example, the SF-36 in Appendix C.2).

Assessment of MIDs

In general, the comments contained few concrete suggestions on how to address the variability of MIDs and the existing difficulties in evaluating currently available studies on MIDs. In 2020, Devji et al. published an instrument for assessing the informative value of anchor-based definitions of MIDs [5]. In this paper, a group of researchers from McMaster University describes the development of quality criteria for studies for determining MIDs and the results of a study on the reliability of the newly developed instrument. They focus solely on anchor-based methods; distribution-based methods are considered ineffective per se and deemed to neglect the patient's perspective. The quality criteria identified by the authors of the study for assessing the informative value of an MID are understandable in light of the discussion over the past few years. The instrument consists of 5 core criteria (completion of the questionnaire for which the MID is to be determined and of the anchor by patients [or, if necessary, a representative]; ease of understanding and relevance of the anchor for patients; a good correlation between the questionnaire and the anchor; precision of the MID; the anchor's threshold for determining the MID indicates a small but relevant difference) and 4 additional criteria for MIDs determined by assessing changes in the anchor (transition ratings) (optimal time interval between measurements; satisfactory correlation between the questionnaire and the anchor [transition item] at the follow-up measurement; correlation of the anchor [transition item] with the baseline value of the questionnaire; higher correlation between change in the anchor [transition item] and change in the questionnaire than between change in the anchor and the value of the questionnaire at follow-up). These supplementary criteria are relevant because a large proportion of MIDs are determined by assessing changes in the questionnaire and anchor.

Poor publication quality of MID studies

The results of the reliability study [5] are relevant to the discussion regarding IQWiG's proposed methods. To assess the reliability of the newly developed instrument for assessing the validity of anchor-based definitions of MIDs, the group used a random sample from a database of empirically derived MIDs. In this sample, the authors were unable to assess the reliability of the criteria for determining MIDs via queries regarding changes to the anchor, because the necessary information was reported in only a very minor proportion of the studies used to determine the MIDs. The instrument was also used by the same working group in a systematic review of MIDs for PROs. Here, too, it was found that essential quality criteria for MIDs were not reported in the included studies [4]. These findings are consistent with the experience from dossier assessments that evaluate the added benefit of new drugs. In these assessments, the information required to evaluate the informative value of an MID is also often incomplete. The immediate use of this new instrument in dossier assessments would

therefore likely result in a high proportion of negative assessments of the validity of MIDs due to missing information on quality criteria, with the consequence that the submitted analyses would not be used for the assessment. From this perspective as well, the proposal for a generic response threshold relative to the scale range appears to be a sensible step, at least for the period until the quality of studies for determining MIDs has improved.

6.2 International reception of the 15% approach

Meanwhile, IQWiG's specification of a generic response threshold for responder analyses is also attracting international attention. To ensure consistency in the interpretation of results from various PRO instruments and due to the lack of published MIDs, the study by Rea et al. conducted analyses based on a response threshold of 15% of the scale range [35]. Due to the lack of evidence-based thresholds, the Common Sense Oncology initiative and the EORTC recommend a pragmatic approach such as setting a 15% response criterion. To address uncertainties regarding the general applicability of this criterion, the authors recommend sensitivity analyses [36].

7 Conclusion

When analysing PROs measured using complex scales, responder analyses are often conducted based on a response criterion in terms of an individual MID. An MID is defined as the smallest difference in score that is perceptible to a patient and is rated as an improvement or deterioration. However, determining an MID is associated with methodological problems, and for individual instruments, a multitude of MIDs are often published.

The aim of the analysis was therefore to identify a generic response threshold for responder analyses for complex scales – particularly those for measuring PROs – that represents a change perceptible to patients with sufficient certainty.

- An examination of systematic reviews for identifying MIDs for PROs across various therapeutic indications confirmed the aforementioned issue:
  - When determining MIDs, essential methodological quality criteria are often not adhered to. Consequently, a large proportion of the studies included in the systematic reviews could not be used for the extraction of MIDs.
  - Even the MIDs empirically derived from studies conducted with appropriate methods exhibit considerable variability across all PRO instruments.
- However, a systematic examination of the extracted MIDs shows that the majority of empirically derived MIDs lie above the generic threshold of 10% discussed in the literature, but mostly below 20% of the respective scale range. Based on these results, a threshold of 15% of the scale range was therefore considered appropriate to represent a change perceptible to patients with sufficient certainty.
- This is a pragmatic threshold setting based on the MIDs identified in the systematic reviews and not a value calculated from numerical results. This threshold setting is explicitly not a response criterion in terms of an individual MID, but rather a generically applicable response threshold for PRO instruments.
- The General Methods Version 6 [1] further stipulate that analyses with a response threshold of 15% of the scale range are also accepted if they are conducted post hoc (i.e., not predefined in the study protocol of the respective study). This has provided all stakeholders with certainty about which analyses are accepted, particularly in the context of the early benefit assessment of drugs.

8 References

1. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden: Version 6.0 [online]. 2020 [Accessed: 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.

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 [Accessed: 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 [Accessed: 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 [Accessed: 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 [Accessed: 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 [Accessed: 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 [Accessed: 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.

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" [unpublished]. 2022.

40. Bjorner JB. Score Range for the SF-36v2 & SF-12v2 Health Surveys; Standard & Acute Versions [online]. 2022 [Accessed: 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 [Accessed: 09.02.2026]. URL: https://www.healthmeasures.net/explore-measurement-systems/promis.

42. Peipert J, Cella D. Persönliche Kommunikation zu "Scale range of PROMIS measures" [unpublished]. 2023.

The original German version of the working paper is published under https://www.iqwig.de/projekte/ga25-04.html

Appendix A Bibliographic search

1. MEDLINE

Search interface: Ovid

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

The following filters were applied:

- Systematic review: Wong et al. [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 significant*) 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 (experimental animals/not humans)
11	and/5,10
12	11 and (English or German).lg.
13	..l/ 12 yr=2016-Current

Appendix B Tables on the identified MIDs in the individual systematic reviews

Table 3: Summary of the MIDs from the studies included in the Doganay review [19]

Instrument	Scale range	% range of MIDs[a] (number of studies)
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. Percentage of the respective MID relative to the instrument's scale range.

MID: minimally important difference

Table 4: Summary of the MIDs from the studies included in the St. Pierre review [28]

Instrument	**Scale range**	**% range of MIDs[a] (number of studies)**
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. Percentage of the respective MID relative to the instrument's scale range. MID: minimally important difference		

Table 5: Summary of the MIDs from the studies included in the Hao review [22]

Instrument	**Scale range**	**% range of MIDs[a] (number of studies)**
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. Percentage of the respective MID relative to the instrument's scale range. MID: minimally important difference		

Table 6: Summary of the MIDs from the studies included in the Ebrahim review [20]

Instrument	**Scale range**	**% range of MIDs[a] (number of studies)**
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 the 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. Percentage of the respective MID relative to the instrument's scale range. b. Different ranges for subscales. MID: minimally important difference		

Table 7: Summary of the MIDs from the studies included in the Nordin review [25]

Instrument	**Range of the scale**	**% range of MIDs[a] (number of studies)**
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. Percentage of the respective MID relative to the instrument's scale range. MID: minimally important difference		

Table 8: Summary of the MIDs from the studies included in the Jayadevappa review [23]

Instrument	Scale range	% range of MIDs[a] (number of studies)
Cardiovascular diseases		
Barthel Index	0–20	9 (1)
Stroke Rehabilitation Assessment of Movement (STREAM)[b]	0–20 (0–30)	10–16 (1)
Urogenital tract disorders		
Incontinence Quality of Life (I-QOL)	0–100	6–7 (1)
Qualiveen-30	0–4	9–18 (1)
Respiratory diseases		
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)
Paediatrics		
Childhood Health Assessment Questionnaire (CHAQ)[b]	0–24 (0–3)	< 1–8 (1)
Acute Otitis Media Symptom Severity (AOM-SOS)	0–10	38 (1)
Various		
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)
Musculoskeletal disorders		
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)
Oncology		
European Organization 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)

Table 8: Summary of the MIDs from the studies included in the Jayadevappa review [23]

Instrument	**Scale range**	**% range of MIDs[a] (number of studies)**
Immune Thrombocytopenic Purpura Patient Assessment Questionnaire (ITP-PAQ)	0–100	3–16 (1)
a. Percentage of the respective MID relative to the instrument's scale range. b. Different ranges for total score and subscales. MID: minimally important difference		

Table 9: Summary of the MIDs from the studies included in the Alma review [13]

Instrument	**Range of the scale**	**% range of MIDs[a] (number of studies)**
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. Percentage of the respective MID relative to the instrument's scale range. MID: minimally important difference		

Table 10: Summary of the MIDs from the studies included in the Ousmen review [27]

Instrument	**Scale range**	**% range of MIDs[a] (number of studies)**
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. Percentage of the respective MID relative to the instrument's scale range. MID: minimally important difference		

Appendix C Specific instruments

For the following instruments, specific guidelines for implementing the 15% approach were published as part of dossier assessments and are presented below.

C.1 EORTC QLQ-C30 and its supplementary modules: Suitability of the 10-point response thresholds for analyses of the EORTC modules

The EORTC QLQ-C30 uses 30 items to assess both health-related quality of life and general symptoms in cancer patients. The instrument consists of a scale for global health status, 5 functional scales, and 8 symptom scales. Items are rated on a scale of 1 to 4, and both questions regarding global health status are rated on a scale of 1 to 7. For analysis, the scales are converted to values ranging from 0 to 100.

The various scales each contain a different number of items. The functional scales include physical function (5 items), role function (2 items), emotional function (4 items), cognitive function (2 items), and social function (2 items). The scale for global health status includes 2 items. The symptom scales include fatigue (3 items), pain (2 items), nausea and vomiting (2 items), dyspnoea, insomnia, loss of appetite, constipation, and diarrhoea (1 item each).

For the scales of the EORTC-QLQ C30, based in part on the work of Osoba [32], an international MID of 10 points was determined for all scales, each corresponding to 10% of the scale range. Against this background, it was examined whether, based on the construction and number of items of the respective scales, differences in the results arise when using a response criterion of 10 points compared to a response criterion of 15% of the scale range of the EORTC scales (corresponding to 15 points).

Table 11 provides an overview of the ranges of the raw scores and the smallest possible score increments for the individual scales depending on the number of items per scale. For each scale, the smallest possible score increment determines whether analyses using the response criteria of 10 or 15 points yield different results.

Table 11: Relevance of the different response thresholds (10 points / 15 points) for the scales of the EORTC QLQ-C30

EORTC QLQ-C30 Scale (number of items, range of values)[a]	**Range per scale [raw values][b]**	**Number of possible score increments on the scale[c] [n]**	**Smallest possible score increment on a standardized scale (0–100)[d] [points]**	**Actual change in response ≥ 10 points[e] [points]**	**Actual change in response ≥ 15 points[f] [points]**
Dyspnoea, insomnia, loss of appetite, constipation, diarrhoea (1 item each, score range 1–4)	3	3	33.3	33.3	33.3
Role function, cognitive function, social function, pain, nausea, and vomiting (2 items each, scale 1–4)	3	6	16.7	16.7	16.7
Global health status (2 items, scale 1–7)	6	12	8.3	16.7	16.7
Fatigue (3 items, scale 1–4)	3	9	11.1	11.1	**22.2**
Emotional function (4 items, scale 1–4)	3	12	8.3	16.7	16.7
Physical function (5 items, scale 1–4)	3	15	6.7	13.3	**20.0**

a. The "financial difficulties" scale (1 item) is not included in the benefit assessment.
b. Each item can take values from 1 to 4 (or from 1 to 7 for items on the global health status scale). The raw score for the scale is calculated as the sum of the individual items divided by the number of individual items.
c. The number of possible score increments on the scale is calculated by multiplying the number of possible score increments per item by the number of items.
d. The smallest possible score increment on the standardized scale is calculated by dividing the range of the standardized scale (0–100) by the number of possible score increments on the scale.
e. Actual change on the scale when applying a response threshold of at least 10 points.
f. Actual change on the scale when applying a response threshold of at least 15 points.

This means that, when applying the response thresholds of 10 and 15 points in a responder analysis, different numbers of responders may result for only 2 of the 14 scales of the EORTC QLQ-C30, while they lead to identical numbers of responders for the remaining scales. For example, for the 5 scales that contain only 1 item, the smallest possible change in the scale is a change of 33.3 points on the 0-to-100 standardized scale. As a result, these patients are included as responders in the analysis in both an analysis with a response threshold of 10 points and an analysis with a response threshold of 15 points. The number of responders is thus identical in both analyses. This applies equally to the 6 scales with 2 items (smallest

possible score increment of 16.7 points) as well as to the scale with 4 items (score increment of 16.7 points; the smallest possible increment of 8.3 points does not meet either response threshold). Differences arise only for the scales on fatigue (3 items) and physical function (5 items): For the 3-item scale, a response threshold of 10 points corresponds to a score increment of 1 item on the scale (11.1 points), while a response threshold of 15 points requires 2 score increments (22.2 points). Similarly, the corresponding values for the 5-item scale are 2 score increments for the 10-point threshold (13.3 points) and 3 for the 15-point threshold (20.0 points). Overall, therefore, for 12 (86%) of 14 statements made based on responder analyses of the EORTC QLQ-C30 scales, both response thresholds yield identical results. Differences may arise only for the 2 scales mentioned. However, the difference between the 2 response thresholds is only 1 increment on the respective scale. In this specific situation of the EORTC, therefore, the analysis using a 10-point response threshold is considered a sufficient approximation of an analysis using a 15% threshold (15 points) and is used for the benefit assessment.

The analysis presented above can also be applied to the indication-specific add-on modules of the EORTC, which are used exclusively in combination with the EORTC QLQ-C30. For example, the add-on module for hepatocellular carcinoma (HCC18) comprises a total of 8 scales, 6 of which contain 1 or 2 items each and thus yield identical results for both response thresholds. Only 2 scales (fatigue, 3 items, and nutrition, 5 items) differ – as shown above – by 1 score increment on the scale when applying the 2 response thresholds.

Overall, more than 70% of the statements are thus identical, and for the scales that differ, the difference never exceeds 1 score increment on the scale. Thus, for the combined use of the EORTC QLQ-C30 with the HCC18, the analysis using a response threshold of 10 points is considered a sufficient approximation of an analysis using a 15% threshold and is used for the benefit assessment.

This approach can also be applied to the majority of the EORTC's indication-specific add-on modules, which are used exclusively in combination with the EORTC QLQ-C30 (see also [11]).

C.2 Derivation of the response criteria for the Physical Component Summary (PCS) and Mental Component Summary (MCS) of the SF-36

The total scores of the SF-36 (Physical Component Summary [PCS] and Mental Component Summary [MCS]) are, after transformation, standardized values that no longer have a scale range of 0 to 100. How the scale range can be determined in this specific situation involving standardized values is explained below. To this end, the calculation of the PCS and the MCS is first outlined (see Chapter 5 of the SF-36 Manual [38]), and then a scale range is specified as the basis for determining a response criterion. All information applies to Version 2 of the SF-36 with a recall period of 4 weeks.

The SF-36 consists of 36 items, 35 of which are assigned to the 8 domains listed in Table 11. Domain scores are calculated by summing the item scores for each domain and then transforming the totals to values between ≥ 0 and ≤ 100. The domain scores for each participant are then transformed into a z-score based on the means and standard deviations of the 2009 SF-36 normative sample (see Chapter 14 in the manual [38]). The z-scores are then converted into t-scores by multiplying them by a factor of 10 and adding 50. This standardization results in a distribution with a mean of 50 and a standard deviation of 10 for the general population (represented by the SF-36 normative sample). A T-score below 50 indicates poorer function or a worse condition compared to the respective mean value of the general population; for a T-score above 50, the opposite is true. These calculations result in the scale range of the 8 domains no longer being 100.

To calculate both higher-order component summaries (PCS and MCS), the z-scores of the 8 domains are each weighted by the coefficients listed in Table 11 and then summed. The z-scores of both component summaries are subsequently converted into T-scores, just like the z-scores of the domains, by multiplying them by a factor of 10 and adding 50. As with the 8 domain scores, this standardization results in a distribution with a mean of 50 and a standard deviation of 10 for a general population (represented by the SF-36 norm sample). A T-score below 50 indicates a lower health-related quality of life compared to the average health-related quality of life of the general population; for a T-score above 50, the opposite is true.

Table 12: Number of items per domain and weighting of the domains in the PCS and MCS

	Coefficients	
Domain (number of items)[a]	**PCS**	**MCS**
Physical functioning (10)	0.42402	−0.22999
Physical role function (4)	0.35119	−0.12329
Physical pain (2)	0.31754	−0.09731
General perception of health (5)	0.24954	−0.01571
Vitality (4)	0.02877	0.23534
Social functioning (2)	−0.00753	0.26876
Emotional role functioning (3)	−0.19206	0.43407
Psychological well-being (5)	−0.22069	0.48581
a. Only 35 of the 36 items are used to calculate the PCS and the MCS. MCS: Mental Component Summary; PCS: Physical Component Summary		

As shown in Table 12, the coefficients for each domain and component summary (PCS, MCS) are either positive or negative. The inclusion of negative coefficients means that the minima and maxima of the PCS and MCS cannot be calculated by assigning the best or worst value to each item (Scenario 1 in Table 13). If, on the other hand, the scale range is calculated by

selecting items in such a way that the PCS or MCS is maximized or minimized, a larger scale range results (Table 13, Scenario 2 calculated using the 1998 normative sample and Scenario 3 calculated using the 2009 normative sample [Personal communication with Jakob Bue Bjørner, MD, PhD, Chief Science Officer at QualityMetric [39]] [40]). In addition to the calculated possible scale ranges (Scenarios 1, 2, and 3), the SF-36 manual [38] provides the empirical minima and maxima from the 2009 normative sample (Table 13, Scenario 4).

Table 13: Minima and maxima of the SF-36 PCS and MCS in various scenarios

	Scenario 1: PCS and MCS results assuming the lowest and highest possible values per item, respectively (calculated using the 1998 normative sample)[a]		**Scenario 2: Minimization or maximization of the PCS or MCS (calculated using the 1998 normative sample)[a]**		**Scenario 3: Minimization or maximization of the PCS or MCS (calculated using the 2009 normative sample)[b]**		**Scenario 4: Observed values in the 2009 normative sample[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. Authors' own calculation.
b. Personal communication with Jakob Bue Bjørner, MD, PhD, Chief Science Officer at QualityMetric [39,40].
c. Information in Table 7.1 of the SF-36 Manual [38].

Max: maximum; MCS: Mental Component Summary; Min: Minimum; PCS: Physical Component Summary

Of the theoretical minima and maxima determined in Table 13 (Scenarios 1, 2, and 3), the values according to Scenario 1 are not suitable as a basis for determining a response criterion, as they do not cover the maximum possible scale range. Although the minima and maxima according to Scenarios 2 and 3 cover the maximum possible scale range, the minima and maxima occur only when a maximum mental impairment is accompanied by a minimum physical impairment (and vice versa) due to the weighting of the components of the total scores.

Due to this algorithm of the analysis, the scale range of the empirical minima and maxima from the 2009 normative sample (Scenario 4) is used for both component summaries of the SF-36 to determine the response criterion (15% of the scale range). With a minimum of approximately 7 (PCS) or 6 (MCS) and a maximum of approximately 70 for both domains (see Table 13), this results in a scale range of approximately 63 (PCS) or 64 (MCS), resulting in a response criterion of just under 10 points for both component summaries (PCS: 9.4; MCS: 9.6) at 15% of the scale range. Thus, 10 points for the PCS and MCS of the SF-36 each represent a suitable approximation for a response criterion using 15% of the scale range.

For the SF-36 with a recall period of 1 week, the following minima and maxima are given in the manual (Table 7.1) [38]: PCS: 10.8 and 75.5; MCS: 5.6 and 69.7. This results in scale ranges of approximately 64.7 and 64.1 and response criteria of 9.7 and 9.6. Here, too, 10 points for the PCS and MCS of the SF-36 each represent a suitable approximation for a response criterion using 15% of the scale range (see also [12]).

C.3 PROMIS / Item Banks

PROMIS item banks are administered via computer-adaptive testing (CAT) [41]. In a CAT, the number and difficulty level of the items are not predefined but are determined by the patients' responses. Additionally, raw scores are converted into T-scores, which (at least theoretically) range from minus infinity to infinity.

For data collected using the PROMIS system, a proposal for a response criterion was discussed with David Cella, Devin Peipert, and colleagues [42]. Based on the scale distribution data from HealthMeasures, a functional scale range of 20 to 80 can be assumed, since very few patients have T-scores below 20 or above 80. Based on this scale range, a threshold of 9 T-score points corresponds to 15% of the scale span.