Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study

Tahira Devji,[1] Alonso Carrasco-Labra,[1] Anila Qasim,[1] Mark Phillips,[1] Bradley C Johnston,[1,2] Niveditha Devasenapathy,[3] Dena Zeraatkar,[1] Meha Bhatt,[1] Xuejing Jin,[4] Romina Brignardello-Petersen,[1] Olivia Urquhart,[5] Farid Foroutan,[1] Stefan Schandelmaier,[1] Hector Pardo-Hernandez,[6,7] Robin WM Vernooij,[8] Hsiaomin Huang,[9] Yamna Rizwan,[10] Reed Siemieniuk,[1] Lyubov Lytvyn,[1] Donald L Patrick,[11] Shanil Ebrahim,[1] Toshi Furukawa,[12] Gihad Nesrallah,[13,14,15] Holger J Schünemann,[1,16] Mohit Bhandari,[1,17] Lehana Thabane,[1] Gordon H Guyatt[1,16]

For numbered affiliations see end of the article

Correspondence to: T Devji devjits@mcmaster.ca (or @TahiraDevji on Twitter (ORCID 0000-0001-8414-7410)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;369:m1714 http://dx.doi.org/10.1136/bmj.m1714

Accepted: 31 March 2020

ABSTRACT

OBJECTIVE

To develop an instrument to evaluate the credibility of anchor based minimal important differences (MIDs) for outcome measures reported by patients, and to assess the reliability of the instrument.

DESIGN

Instrument development and reliability study.

DATA SOURCES

Initial criteria were developed for evaluating the credibility of anchor based MIDs based on a literature review (Medline, Embase, CINAHL, and PsycInfo databases) and the experience of the authors in the methodology for estimation of MIDs. Iterative discussions by the team and pilot testing with experts and potential users facilitated the development of the final instrument.

PARTICIPANTS

With the newly developed instrument, pairs of masters, doctoral, or postdoctoral students with a background in health research methodology independently evaluated the credibility of a sample of MID estimates.

MAIN OUTCOME MEASURES

Core credibility criteria applicable to all anchor types, additional criteria for transition rating anchors, and inter-rater reliability coefficients were determined.

RESULTS

The credibility instrument has five core criteria: the anchor is rated by the patient; the anchor is interpretable and relevant to the patient; the MID estimate is precise; the correlation between the anchor and the outcome measure reported by the patient is satisfactory; and the authors select a threshold on the anchor that reflects a small but important difference. The additional criteria for transition rating anchors are: the time elapsed between baseline and follow-up measurement for estimation of the MID is optimal; and the correlations of the transition rating with the baseline, follow-up, and change score in the patient reported outcome measures are satisfactory. Inter-rater reliability coefficients (κ) for the core criteria and for one item from the additional criteria ranged from 0.70 to 0.94. Reporting issues prevented the evaluation of the reliability of the three other additional criteria for the transition rating anchors.

CONCLUSIONS

Researchers, clinicians, and healthcare policy decision makers can consider using this instrument to evaluate the design, conduct, and analysis of studies estimating anchor based minimal important differences.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Interpreting results from patient reported outcome measures is critical for healthcare decision making

The minimal important difference, a measure of the smallest change in a measure that patients consider important, can greatly facilitate judgments of the magnitude of effect on patient reported outcomes

The credibility of minimal important difference estimates varies, and guidance on determining credibility is limited

WHAT THIS STUDY ADDS

An instrument to evaluate the design, conduct, and analysis of studies measuring minimal important differences has been developed

This instrument aims to allow users to distinguish between unreliable and credible minimal important differences to optimise the presentation and interpretation of results from patient reported outcome measures in clinical trials, systematic reviews, health technology assessments, and clinical practice guidelines

This instrument will also aim to promote higher standards in methodology for robust anchor based estimation of minimal important differences

Introduction

The role of the patient's perspective in clinical research has increased over the past 30 years. Questionnaires looking at health status from the patient's perspective—patient reported outcome measures—is an important strategy in determining the effect of interventions. Despite improvements in establishing their validity, reliability, and responsiveness, interpretation of outcome measures reported by patients is challenging.

Interpretability deals with how to determine differences in scores for patient reported outcome measures that constitute trivial, small but important, moderate, or large differences.[1 2] To help in the design and interpretation of trials evaluating the effect of an intervention on patient reported outcomes,

researchers developed a concept called the minimal important difference (MID).[3 4] The MID provides a measure of the smallest change in an outcome measure that patients perceive as an important improvement or deterioration,[3 4] and can be used as a reference point for judging the magnitude of treatment effects in clinical trials and systematic reviews.

The widely accepted optimal approach to establishing an MID for a patient reported outcome measure relates a score on the instrument to an independent measure—an external criterion or anchor—that is understandable and relevant to patients.[5] The most widely used anchor is the patient's global rating of change, also referred to as a transition rating. An example of a typical transition rating question would be "Since last month when we started the new treatment, are you feeling better or worse and, if so, to what extent?", with responses of no change, small but important, moderate or large improvement, or worsening. Other anchors include measures of satisfaction, occurrence of an event, or other patient reported outcome measures assessing health status.

A second, but much less effective, approach to estimating the MID involves distribution based methods. These methods rely solely on the statistical characteristics of the study sample (eg, 0.5 standard deviation of scores for patient reported outcome measures) and fail to incorporate the patient's perspective.[6 7]

The methodology behind an anchor based MID relies on two key components: choice of anchor and statistical method to estimate the MID. Some of these choices are more satisfactory than others; poor choices can lead to MIDs that mislead, and misleading MIDs will result in seriously flawed interpretation of results for patient reported outcome measures in clinical trials and systematic reviews. For MIDs to help inform patient care, investigators and decision makers (including those performing clinical trials, authors of systematic reviews, developers of clinical practice guidelines, and regulatory authorities, and their audiences of clinicians and patients) must be able to distinguish between unreliable and credible or trustworthy MIDs.

How likely is the design and conduct of studies measuring MIDs to have provided robust estimates? Currently, no accepted standards exist for evaluating the credibility of an anchor based MID. Here, we describe the development of an instrument to evaluate the credibility of anchor based MIDs and report the inter-rater reliability of the instrument.

Methods

Development of a credibility instrument

A steering committee was set up that included clinicians, health research methodologists, clinical epidemiologists, and a psychiatrist (TD, AC-L, TF, BCJ, GN, DLP, and GHG), with substantial experience in measuring health status. The steering committee coordinated the development of the credibility instrument, recruited collaborators, prepared and revised documents, and provided administrative support.

Selection and development of candidate credibility criteria

Our research group conducted a systematic review to develop an inventory of anchor based MIDs for patient reported outcome measures (A Carrasco-Labra, personal communication, 2020).[8] To develop criteria for assessing the credibility of anchor based MIDs, during study selection for the MID inventory, we simultaneously screened for articles reporting on key issues or considerations about anchor based methods. Specifically, we selected articles with theoretical descriptions, summaries, commentaries, and critiques suggesting one or more criteria for the credibility of any aspect of anchor based methodology for estimation of MIDs. We searched Medline, Embase, CINAHL, and PsycInfo from 1989 to April 2015, to identify relevant articles for both projects. The search strategy, adapted to each database, included terms representing the MID concept and terms looking at patient reported outcome measures (appendix 1).

We used a standardised data extraction form to abstract candidate criteria for establishing the credibility of an anchor based MID from the methods articles selected (appendix 4). We also extracted excerpts for any rationale or explanation provided by authors for why a specific criterion would increase or decrease credibility. After data extraction, through qualitative analysis, we developed a taxonomy with a deductive approach and categorised criteria according to themes.[9 10]

The steering committee reviewed and discussed the results of the coded data extraction, and evaluated the themes that emerged from the qualitative analysis. Findings from the survey of the literature, coupled with our groups' experience with methods of establishing MIDs,[1 2 5 6 11-28] allowed for full discussion of key credibility concepts. Issues that arose based on our experience were the effect of varying correlations between anchor and target instrument, the effect of duration of time required for recall, the relation between sample size and precision of MID estimates, and the relative merits of alternative statistical approaches for estimation of MIDs. The steering committee reviewed the candidate criteria and evaluated the importance of each. Criteria were eliminated when redundancy or overlap existed, and when criteria were not optimally relevant. The steering committee drafted an initial version of the instrument, including clearly worded items, associated response options for each item, and instructions for completing each item.

Piloting and user feedback

We conducted an iterative process of pilot testing and user feedback. We presented the initial instrument to a convenience sample of experts (about seven health research methodologists and clinical epidemiologists with expertise in instrument development, MID estimation, and patient reported outcomes) and target users (about two clinicians, 13 authors of systematic reviews, and three guideline developers). These individuals reviewed the clarity,

wording, comprehensiveness, and relevance of the items of the instrument, and provided suggestions for the instrument. We incorporated this feedback. Based on subsequent work, including application of the draft instrument to anchor based MID estimation studies in our MID inventory[8] and more applications of the instrument to inform the development of a clinical practice guideline,[29] we modified the instrument and reduced the number of items. This process continued until the steering group reached consensus that the instrument would prove optimal for use.

Reliability study of the credibility instrument

Sample of MID estimates and raters

In our previously mentioned inventory of anchor based MIDs, we summarised more than 3000 estimates and their associated credibility, including MIDs for patient reported outcome measures across different populations, conditions, and interventions, obtained with different anchors and statistical methods.[8] We enlisted help from masters, doctoral, and postdoctoral students with a background in health research methodology to conduct study screening, data extraction, and the credibility assessment. Before starting the review process, the reviewers received extensive training on the methodology of MIDs, including background reading of key methods articles on MIDs, web teleconferences to review screening and data extraction materials, and pilot and calibration exercises. Teams of two reviewers independently extracted relevant data from the studies selected for each MID estimate, collecting information on study design, characteristics of the patient reported outcome measure, anchor and analytical method, sample size, the MID estimate and associated measure of precision, and time elapsed between administration and follow-up assessments of the patient reported outcome measure and anchor (for longitudinal designs). The reviewers applied the newly developed instrument to evaluate the credibility of the MID estimates.

Sampling method

For a random sample of 200 MID estimates from our inventory, we retrieved the credibility assessments performed by each pair of reviewers with the newly developed instrument. We sampled in excess (see sample size below) to account for potential discrepancies in the MIDs extracted between reviewers and incomplete data (eg, where one reviewer might have missed an MID reported in the study, we would only have one credibility assessment). Because the questions in the extension of the credibility instrument apply only to MIDs estimated with transition rating anchors, and only 50% of the initial sample of 200 MIDs used transition anchors, we randomly sampled an additional 50 MID estimates to meet the required sample size for the relevant reliability analyses. To ensure observations in our sample were independent of each other, when one study reported multiple MIDs, we included only the first MID estimate extracted for that study.

Sample size

We tested the reliability of our credibility instrument with classical test theory.[30] Because assessments of credibility involve subjective judgments and different individuals collecting data might experience and interpret phenomena of interest differently, we measured inter-rater reliability. According to Shoukri,[31] considering two raters per MID estimate, an expected reliability of 0.7, with a desired 95% confidence interval width of 0.2, and an α of 0.05, would require a minimum of 101 MIDs assessed per rater.

Analysis

For each item of the instrument, we calculated inter-rater reliability and the associated 95% confidence interval, measured by a weighted kappa, κ, with quadratic weights assigned by the formula: $w_i=1-(i^2/(k-1)^2)$, where i is the difference between categories (response options) and k is the total number of categories. The use of quadratic weights implies that response options for the credibility criteria are ordinal and equidistant. In the absence of information in the primary study to make an informed judgment, the "impossible to tell" response option can be used (see credibility instrument in the results section below). We consider that this rating reflects low certainty in terms of credibility and thus we combined these responses with ratings of "definitely no" in the reliability analysis. We considered a reliability coefficient of at least 0.7 to represent good inter-rater reliability.[32-34]

Patient and public involvement

Patients and the public were not involved in the design, conduct, or reporting in this methodological research, as our instrument is a critical appraisal tool that is intended for researchers and decision makers who require MIDs for interpretation of patient reported outcome measures, including clinical trial investigators, authors of systematic reviews, guideline developers, clinicians, funders, and policy makers.

Results

We identified 41 relevant articles on MID methods[4-6 14 22 27 28 35-68] that informed the item generation stage of the development of the instrument (fig 1). There were two major modifications from the first draft[69] to the final instrument. In the first, we removed three items (items 2, 4, and 6 of the first draft) because of issues of redundancy and relevance; rephrased one item dealing with to what extent the anchor and the patient reported outcome measure are measuring the same construct (item 5 of the first draft); and added one new item looking at the precision around the point estimate of the MID. In the second modification, we added a new item evaluating whether the anchor threshold selected for estimation of the MID reflected a small but important difference, and

developed more criteria for assessing the credibility of a transition rating anchor (described below).

Credibility instrument

The instrument has five criteria essential for determining the credibility of any anchor based MID (table 1). In our inventory of anchor based MIDs[8] and a separate systematic review to identify MIDs for knee specific patient reported outcome measures,[29] we found that MIDs were most often derived with transition rating anchors. Transition rating anchors require patients to recall a previous health state and compare it with how they are feeling now. This retrospection required criteria to ensure that transition ratings accurately reflect the change in health status and are not unduly influenced by the baseline or endpoint status; thus, for this situation, we developed a four item extension of the core credibility instrument (table 1). Below, we describe each question in the instrument with an explanation of the relevance of the item for evaluating credibility (the full version of the instrument is in appendix 2 with three worked examples in appendix 3, where we have applied our instrument to assess the credibility of three MID estimates, each from a published study).

Except for the first item, which has a yes or no response, each item has a five point adjectival scale. The response options for items in the instrument are: definitely yes; to a great extent; not so much; definitely no; and impossible to tell. A response of definitely yes indicates no concern about the credibility of the MID estimate. Responses of definitely yes and definitely no imply that information provided in the MID report under evaluation allows an unequivocal judgment in relation to the item; the responses "to a great extent" and "not so much" indicate less certainty. In the absence of information or sufficient detail to make an informed judgment about credibility, the response option "impossible to tell" can be used.

Explanation of the core credibility items

Item 1: Is the patient or necessary proxy responding directly to both the patient reported outcome measure and the anchor?

An anchor based method for estimating an MID involves linking a specific patient reported outcome measure (eg, short form 36, Beck depression inventory, chronic respiratory questionnaire) to an external criterion, such as a patient or physician transition rating, another patient reported outcome measure, or a clinical endpoint (eg, concentration of haemoglobin, Eastern Cooperative Oncology Group performance status). Patient reported anchors are more desirable than clinical measures or those that are assessed by a clinician. Situations where the patient cannot directly provide information to inform the outcome (eg, elderly individuals with dementia, infants, and pre-verbal toddlers) require a proxy respondent. We suggest using the same standards recommended for a patient directly responding to the outcome measure when evaluating the credibility of MIDs for a necessary proxy reported outcome measure on behalf on the patient.

Item 2: Is the anchor easily understandable and relevant for patients or necessary proxy?

A suitable anchor is one that is easily understandable and is highly relevant to patients. Typical appropriate anchors are global ratings of change in health status,[19 70-72] status on an important and easily understood measure of function,[73] the presence of symptoms,[74] disease severity,[75] response to

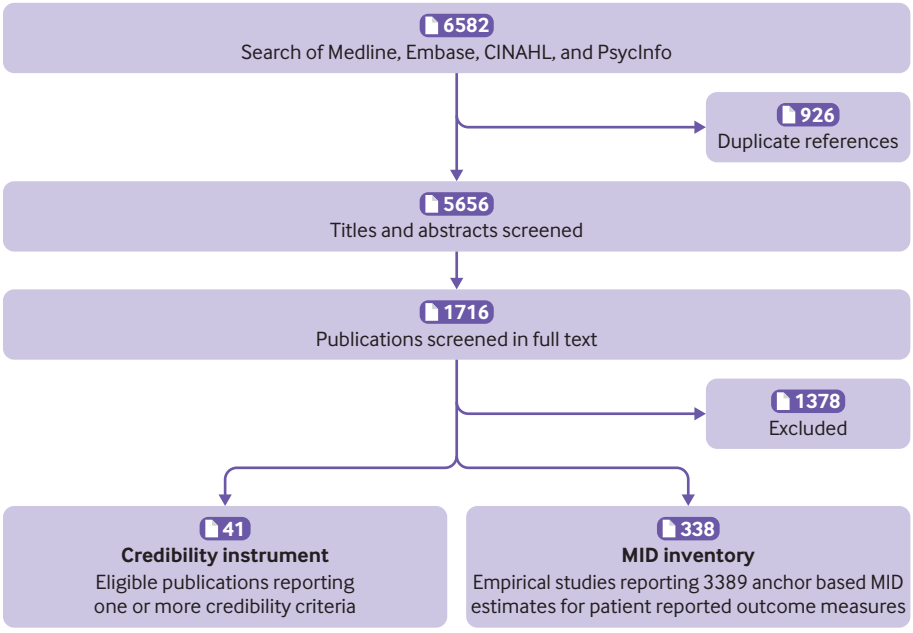

Fig 1 | Selection of studies for the development of the minimal important difference (MID) inventory and the credibility instrument

doi: 10.1136/bmj.m1714 | *BMJ* 2020;369:m1714 | the bmj

Table 1 | Credibility instrument for judging the trustworthiness of minimal important differences

Signalling question	Response options	
	High credibility	Low credibility
Core criteria		
Is the patient or necessary proxy responding directly to both the PROM and the anchor?	Yes	No/impossible to tell
Is the anchor easily understandable and relevant for patients or necessary proxy?	Definitely yes/to a great extent	Definitely no/not so much/impossible to tell
Has the anchor shown good correlation with the PROM?		
Is the MID precise?		
Does the threshold or difference between groups on the anchor used to estimate the MID reflect a small but important difference?		
Additional criteria for transition rating anchors		
Is the amount of elapsed time between baseline and follow-up measurement for MID estimation optimal?	Definitely yes/to a great extent	Definitely no/not so much/impossible to tell
Does the transition item have a satisfactory correlation with the PROM score at follow-up?		
Does the transition item correlate with the PROM score at baseline?		
Is the correlation of the transition item with the PROM change score appreciably greater than the correlation of the transition item with the PROM score at follow-up?		

PROM=patient reported outcome measure; MID=minimal important difference.

treatment,[75 76] or the prognosis for future events, such as death,[74 77 78] use of healthcare facilities,[79] or job loss.[74 80 81]

Item 3: Has the anchor shown good correlation with the patient reported outcome measure?

The usefulness of anchor based approaches is critically dependent on the relation between the patient reported outcome measure and the anchor. When determining the credibility of the MID, we consider how closely the anchor is related to the target patient reported outcome measure and give greater importance to MIDs generated from closely linked concepts; the anchor and patient reported outcome measure should be measuring the same or similar underlying constructs, and therefore should be appreciably correlated. A moderate to high correlation (at least 0.5) suggests the validity of the anchor.[14 82 83] An anchor that has low or no correlation with the patient reported outcome measure will likely give inaccurate MID estimates. The instrument has a guide for judging the correlation coefficient.

Item 4: Is the MID precise?

To judge precision, we focus on the 95% confidence interval around the point estimate of the MID. We provide a guide for judging precision when the investigators report the 95% confidence interval around the MID estimate based on the likelihood that inferences about the magnitude of a treatment effect would differ at the extremes of the confidence interval. When authors do not provide a measure of precision, the number of patients included in the estimation of the MID gives an alternative criterion for judging precision. We provide guidance on appropriate sample size based on the relation between sample size and precision in studies in the inventory that reported 95% confidence intervals.

Item 5: Does the threshold or difference between groups on the anchor used to estimate the MID reflect a small but important difference?

To respond to this credibility question, a judgment is needed on whether the selected threshold or groups compared on the anchor reflect a small (rather than moderate or large) but important difference. Even after the threshold is set, many analytical methods can be used to compute the MID, and whether the chosen method of analysis calculates an MID needs to be determined. Box 1 provides a framework for making these judgments, and box 2 has examples of high and low credibility MIDs estimated with different types of anchors.

Explanation of additional items for transition rating anchors

Item 1: Is the amount of elapsed time between baseline and follow-up measurement for MID estimation optimal?

Despite the intuitive appeal of transition questions, patients have considerable difficulty recalling previous health states,[14 49 87] and the longer the time patients have to remember, the greater the difficulty.[14 49] Patients can often remember previous states for up to four weeks[14]; as time extends into months, patients are more likely to confuse change over time with current status.[49]

Judgments for items 2-4 of the extension for transition rating anchors requires knowledge of the directional characteristics of the patient reported outcome measure and transition scale. In the instrument, we provide guidance to deal with situations where higher scores on both the patient reported outcome measure and anchor represent the same direction (that is, both represent a worse or better condition) and when they represent different directions.

Item 2: Does the transition item have a satisfactory correlation with the score for the patient reported outcome measure at follow-up?

Ideally, the correlation between the transition rating with the score at baseline and the transition rating with the score at follow-up would be equal and opposite, an ideal that seldom occurs. To the extent that the score at follow-up shows at least some correlation with the transition, the MID estimate is more credible than if there were no correlation.[14 68]

Box 1: Considerations for judging whether the minimal important difference represents a small but important difference

1. What is the original scale of the anchor and is it transformed in any way?
2. Does the scale (or transformed scale) of the anchor capture variability in the underlying construct?
3. What is the threshold used or comparison being made on the anchor? Does this threshold or comparison represent a difference that is minimally important?
4. Does the analytical method ensure that the minimal important difference represents a small but important difference? The last example in box 2 shows how a poorly chosen analytical method could give misguided minimal important difference estimates.

Box 2: Examples of high and low credibility ratings for item 5 of the credibility instrument

High credibility

- Investigators calculated the minimal important difference (MID) for the pain domain of the Western Ontario and McMaster University Osteoarthritis Index (WOMAC) as the mean change in the WOMAC pain score in patients who reported themselves as "a little better" to the question "how was the pain in your operated hip during the past week, compared with before the operation," offering response options of extremely better, very much better, much better, better, a little better, a very little better, almost the same or hardly any better, or no change (with parallel responses for worsening).[62]
- To estimate the MID for the National Comprehensive Cancer Network-Functional Assessment of Cancer Therapy (NCCN-FACT) Colorectal Cancer Symptom Index (FCSI-9), investigators compared Eastern Cooperative Oncology Group (ECOG) performance status (score 0-4, higher scores signify worse performance status) at follow-up with baseline performance status. The MID for the FCSI-9 was calculated with the beta coefficients from an analysis of variance model where the dependent variable was the FCSI-9 change score from baseline to week 8 and the independent variable was the ECOG performance status.[84] The investigators decided on a half unit change in the ECOG performance status as a small but important difference—which is assumed to be reasonable—and this threshold was used to derive the MID for the FCSI-9.

Low credibility

- Patients responded to: "Compared to before treatment my back problem is a) much better, b) better, c) unchanged, d) worse." Investigators defined the MID for deterioration for the Oswestry Disability Index by calculating the difference in score between patients who rated themselves as worse and patients who rated themselves as unchanged.[85] This rating has low credibility because worse could mean a little worse or much worse (box 1, framework steps 2 and 3).
- Investigators estimated the MID for the Ability to Perform Physical Activities of Daily Living Questionnaire (APPADL) by taking the difference in mean APPADL change scores for those who achieved 5% or more weight loss from baseline to six months and those who achieved less than 5% weight loss.[86] This rating is problematic because how patients whose weight fell by 6% reacted is not clear—we do not know whether the patients were pleased they had made a substantial weight reduction, had considered the change small but important, or had regarded it as trivial. Also, the researchers used a misguided analytical method. In their group of patients who they classified as having a small but important improvement, they included patients who had a 5%, and also a 10%, 30%, or 50% reduction in weight loss together. Subtracting the APPADL mean change score for the group of patients achieving a less than 5% change in weight loss from those who experienced a change greater than 5% could give an estimate for the MID that constitutes a small, moderate, or large difference, depending on the proportion of patients who achieved large percentage weight losses (box 1, framework step 4). Rather, the use of receiver operating characteristic curve analysis would have been a more appropriate choice.

Item 3: Does the transition item correlate with the patient reported outcome measure score at baseline?

If the score at baseline correlates with the transition rating, we are more confident that patients are taking their baseline status into account when scoring the transition rating.[14 68]

Item 4: Is the correlation of the transition item with the patient reported outcome measure change score appreciably greater than the correlation of the transition item with the patient reported outcome measure score at follow-up?

A correlation of at least 0.5 between the transition rating and the change in patient reported outcome measure is necessary but insufficient to confirm that the transition rating is measuring change, as opposed to current health status. A correlation of the score at follow-up with the transition that is similar or greater than the correlation of the change with the transition indicates that the rating likely reflects only current status, and thus confidence in the MID estimate decreases.[14 68] The instrument provides a guide for judging the correlation coefficients described in items 2-4.

Overall judgment of credibility

Responses to individual items provide the basis for determining an overall judgment of credibility for the MID estimate. We have deliberately avoided a prescriptive approach for reaching an overall judgment and have not scored items, because the relative weights of individual items within the instrument are uncertain and depend on context. Thus the overall credibility judgment for a given MID estimate requires consideration of the severity of the credibility issue for a particular item and the consequence of this issue.

Reliability analyses

The analysis for the assessment of inter-rater reliability included 135 MIDs assessed by two raters for the core credibility criteria and 137 for the first item in the extension criteria. For the remaining items in the extension for transition rating anchors, only 12 studies reported the correlation between the score at follow-up and transition rating described in items 2 and 4, and 10 studies provided the correlation between the score at baseline and transition rating required for item 3. Because of the limited sample sizes, we could not conduct an evaluation of the inter-rater reliability for these items.

Overall, the inter-rater reliability for all items ranged from good (Cohen's κ ≥0.7) to very good (≥0.8) agreement (table 2). The item from the extension criteria looking at duration of follow-up had the highest value for Cohen's κ, and the item on whether the anchor is understandable and relevant, the lowest.

Discussion

Principal findings

We have developed a credibility instrument to evaluate the design, conduct, and analysis of studies measuring

anchor based MIDs. Our instrument is a critical appraisal tool that provides a systematic step-by-step approach to deciding whether a study claiming to establish an MID has trustworthy results. The five criteria in the core credibility instrument proved reliable, with good to excellent agreement between reviewers. The items on whether the anchor is understandable and relevant, and whether the threshold on the anchor represents a small but important difference, had lower, but still satisfactory, inter-rater reliability estimates.

Strengths and limitations of the study

Strengths of the study include the use of the literature and the expertise of the study team in the development of our criteria, and modifications based on expert feedback and extensive experience in applying the instrument. Similar methods have proved successful for developing methodological quality appraisal standards across a wide range of topics.[88-92] We undertook a rigorous assessment that showed the high reliability of the instrument.

Our study has limitations. Although a multidisciplinary team with a broad range of content and methodological expertise led the development of the credibility instrument, these individuals represent only a fraction of the experts in patient reported outcome and MID methodology worldwide. Researchers have not reached consensus on optimal anchor based approaches, types of anchors, and analytical methods, and methodological issues might subsequently emerge that require modification of the instrument.

Reviewers who participated in our reliability study had graduate level methodology training, received extensive instruction on MID methodology, extracted data from at least 30 studies reporting MID estimates, and participated in pilot testing with different iterations of the instrument. Thus reliability might be lower in less well trained and instructed individuals. We have, however, developed detailed instructions and examples (included here and in the appendix) that are likely to enhance reliability in those with less experience than the raters in this study.

We did not conduct a formal evaluation to collect feedback on the usability of our instrument or satisfaction with its use. The instrument did, however, undergo numerous iterations of pretesting, which resolved several issues related to the understanding, comprehensiveness, and overall structure of the instrument.

We could not assess inter-rater reliability for three items in the extension for transition rating anchors, as only 3% of the studies in our inventory of MID estimation studies evaluated the correlations necessary to judge the validity of transition rating anchors. In the future, we anticipate that the availability of this credibility instrument will spur improvements in methodology of the conduct of MID studies. If so, correlations will be regularly reported, and the investigators can look at the reliability of these items.

We have not established the validity of our instrument by formal testing. In other work, however, we have shown that the current criteria for credibility succeed in partially explaining the variability in the magnitude of the MID.[29 93] Our instrument does not deal with the underlying measurement properties of the patient reported outcome measures (that is, validity and responsiveness) and assumes that users will only move forward in evaluating the credibility of MIDs if the instrument has met at least minimal standards of validity and responsiveness.

Implications and future research

Knowing the MID facilitates the interpretation of treatment effects in clinical research, allowing decision makers to determine if patients have had important benefits,[46 94 95] and informing the balance between desirable and undesirable outcomes of interventions. The recent CONSORT PRO Extension (Consolidated Standards of Reporting Trials patient reported outcomes) encourages authors to include discussion of an MID or a responder definition in reports of clinical trials.[14] The demand for increased use of MIDs in trials requires the availability of trustworthy estimates. Since the MID was first introduced over three decades ago,[3 12] methods for calculating the MID have evolved. In our linked inventory of published anchor based MIDs, we identified many statistical methods, each with its own merits and limitations. We also found varying qualities of the anchor, and the threshold selected for defining the MID might not always be optimal. Different methodological and statistical approaches to calculate MIDs will give different estimates for the same patient reported outcome measure.[62 96] Given

Table 2 | Inter-rater reliability coefficients

Item	Weighted κ (95% CI)
Core criteria (n=135 MIDs)	
Is the patient or necessary proxy responding directly to both the PROM and the anchor?	0.80 (0.64 to 0.95)
Is the anchor easily understandable and relevant for patients or necessary proxy?	0.70 (0.66 to 0.76)
Has the anchor shown good correlation with the PROM?	0.90 (0.86 to 0.94)
Is the MID precise?	0.80 (0.67 to 0.87)
Does the threshold or difference between groups on the anchor used to estimate the MID reflect a small but important difference?	0.74 (0.71 to 0.79)
Additional criterion for transition rating anchors (n=137 MIDs)	
Is the amount of elapsed time between baseline and follow-up measurement for MID estimation optimal?	0.94 (0.91 to 0.96)

κ=Cohen's kappa; MID=minimal important difference; PROM=patient reported outcome measure.
The difference in the number of MIDs included in the reliability analysis for the credibility criteria for the transition rating anchors is because of the additional random sample of MIDs retrieved to meet the required sample size (see sampling method section in the methods).

the multiplicity of MID estimates often available for a given patient reported outcome measure and non-standardised methodology, researchers and decision makers in search of MIDs need to critically evaluate the quality of the available estimates.

Flaws in the design and conduct (aspects of credibility) of the studies empirically estimating MIDs can lead to overestimates or underestimates of the true MID. Lack of trustworthy MIDs to guide interpretation of estimates of treatment effects measured by patient reported outcome measures—or worse, availability of misleading MIDs—might result in serious misinterpretations of the results of otherwise well designed clinical trials and meta-analyses. Our credibility instrument provides a comprehensive approach to assessing the credibility of anchor based MIDs. Widespread adoption and implementation of our credibility instrument will facilitate improved appraisal of MIDs by users such as those conducting clinical trials, authors of systematic reviews, guideline developers, clinicians, funders, and policy makers, and also guide the development of trustworthy MIDs.

In developing our inventory of anchor based MIDs, and in other related work,[29 93] we found that the literature often includes a number of candidate MIDs for the same patient reported outcome measure. Moreover, the magnitude of these estimates sometimes varies widely. Several other researcher groups have made similar observations, stressing the importance of improved understanding of factors influencing the magnitude of MIDs.[46 62 97-99] Future research should, therefore, focus on understanding how different methodological and statistical approaches contribute to the variability in MIDs.

Our instrument focuses on the methodological issues that could potentially lead to flawed and thus misleading MIDs, which might in part explain why different methods can give variable estimates. Variability in MIDs, however, can also be related to many other factors, including the clinical setting, patient characteristics (eg, age, sex, disease severity, diagnosis), intervention, and duration of follow-up. Findings from subsequent investigations might provide insights into the appropriate use, in terms of context and trustworthiness, of MIDs for interpretation of patient reported outcome measures in clinical research and practice. For updates to the instrument and associated instructions that may arise from these insights, see www.promid.org.

Conclusions

To better inform management choices, patients, clinicians, and researchers need to know about MIDs to be able to interpret the effects of treatment on patient reported outcome measures. Consideration of the credibility of an MID involves complex judgments. We have developed a reliable instrument that will allow users to distinguish between unreliable and credible MID estimates. This work provides guidance for dealing with the credibility of MIDs to optimise the presentation and interpretation of results from patient reported outcome measures in clinical trials, systematic reviews, health technology assessments, and clinical practice guidelines, and also has important implications for how investigators should conduct future studies on estimating anchor based MIDs.

AUTHOR AFFILIATIONS

[1]Department of Health Research Methods, Evidence and Impact, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4L8, Canada
[2]Department of Community Health and Epidemiology, Dalhousie University, Halifax, NS, Canada
[3]Indian Institute of Public Health, Public Health Foundation of India, Gujarat, India
[4]School of Public Health, University of Alberta, Edmonton, AB, Canada
[5]Center for Evidence Based Dentistry, American Dental Association, Chicago, IL, USA
[6]Iberoamerican Cochrane Centre, Sant Pau Biomedical Research Institute (IIB Sant Pau), Barcelona, Spain
[7]CIBER de Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain
[8]Department of Research, Comprehensive Cancer Organisation, Utrecht, Netherlands
[9]Department of Orthopedic Surgery, University of Michigan, Ann Arbor, MI, USA
[10]Department of Molecular and Cellular Biology, University of Guelph, Guelph, ON, Canada
[11]Department of Health Services, University of Washington, Seattle, WA, USA
[12]Department of Health Promotion and Human Behaviour, School of Public Health, Kyoto University Graduate School of Medicine, Kyoto, Japan
[13]Nephrology Program, Humber River Regional Hospital, Toronto, ON, Canada
[14]Division of Nephrology, University of Western Ontario, London, ON, Canada
[15]Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, ON, Canada
[16]Department of Medicine, McMaster University, Hamilton, ON, Canada
[17]Department of Surgery, McMaster University, Hamilton, ON, Canada

Contributors: TD and AC-L are joint first authors. TD, AC-L, GHG, BCJ, GN, and SE conceived the study idea; TD, AC-L, AQ, MP, and GHG led the development of the credibility instrument; TD, AC-L, AQ, MP, ND, DZ, MB, XJ, RB-P, OU, FF, SS, HP-H, RWMV, HH, YR, RS, and LL extracted data and assessed the credibility of MIDs in our inventory for the reliability analyses; TD and AC-L wrote the first draft of the manuscript; TD, AC-L, GHG, AQ, MP, ND, DZ, RB-P, OU, SS, HP-H, RWMV, LL, BCJ, DLP, SE, TF, GN, HJS, MB, and LT interpreted the data analysis and critically revised the manuscript. TD and AC-L are the guarantors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: This project was funded by the Canadian Institutes of Health Research (CIHR), Knowledge Synthesis (grant No KRS138214). The views expressed in this work are those of the authors and not necessarily represent the views of the CIHR or the Canadian government.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: support from the Canadian Institutes of Health Research (CIHR) for the submitted work; TD, AC-L, and GHG have a patent issued for the Credibility instrument for judging the trustworthiness of minimal

important difference estimates, and a patent pending for the Patient Reported Outcome Minimal Important Difference (PROMID) Database; GHG has received other grants outside the submitted work; MB reports personal fees from AgNovos Healthcare, Sanofi Aventis, Stryker, and Pendopharm, and grants from DJ Orthopaedics and Acumed outside the submitted work; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: No additional data available.

TD, AC-L, and GHG affirm that the manuscript is an honest, accurate, and transparent account of the recommendation being reported; that no important aspects of the recommendation have been omitted; and that any discrepancies from the recommendation as planned (and, if relevant, registered) have been explained.

Dissemination to participants and related patient and public communities: We have planned dissemination of the existence of the instrument and its use to relevant patient communities through health and consumer advocacy organisations, such as the Cochrane Task Exchange, Cochrane Consumer Network, the National Patient-Centred Clinical Research Network, the Society for Participatory Medicine, and Consumers United for Evidence-based Health Care.

1 Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;118:622-9. doi:10.7326/0003-4819-118-8-199304150-00009
2 Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *BMJ* 1998;316:690-3. doi:10.1136/bmj.316.7132.690
3 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10:407-15. doi:10.1016/0197-2456(89)90005-6
4 Schünemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from?*Health Serv Res* 2005;40:593-7. doi:10.1111/j.1475-6773.2005.0k375.x
5 Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77:371-83. doi:10.4065/77.4.371
6 Brożek JL, Guyatt GH, Schünemann HJ. How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. *Health Qual Life Outcomes* 2006;4:69. doi:10.1186/1477-7525-4-69
7 McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA* 2014;312:1342-3. doi:10.1001/jama.2014.13128
8 Carrasco-Labra A, Devji T, Qasim A, et al. Minimal important difference estimates for patient-reported outcomes: The MID inventory. *Manuscript in preparation for submission toJAMA* 2020.
9 Bradley EH, Curry LA, Devers KJ. Qualitative data analysis for health services research: developing taxonomy, themes, and theory. *Health Serv Res* 2007;42:1758-72. doi:10.1111/j.1475-6773.2006.00684.x
10 Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. *Int J Qual Methods* 2017;16:1609406917733847.doi:10.1177/1609406917733847
11 Bryant D, Schünemann H, Brozek J, Jaeschke R, Guyatt G. [Patient reported outcomes: general principles of development and interpretability]. *Pol Arch Med Wewn* 2007;117:5-11. doi:10.20452/pamw.103
12 Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 1987;40:171-8. doi:10.1016/0021-9681(87)90069-5
13 Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol* 1989;42:403-8. doi:10.1016/0895-4356(89)90128-5
14 Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. *J Clin Epidemiol* 2002;55:900-8. doi:10.1016/S0895-4356(02)00435-3
15 Jaeschke R, Guyatt GH, Keller J, Singer J. Interpreting changes in quality-of-life score in N of 1 randomized trials. *Control Clin Trials* 1991;12(Suppl):226S-33S. doi:10.1016/S0197-2456(05)80026-1
16 Johnston BC, Patrick DL, Thorlund K, et al. Patient-reported outcomes in meta-analyses-part 2: methods for improving interpretability for decision-makers. *Health Qual Life Outcomes* 2013;11:211. doi:10.1186/1477-7525-11-211
17 Johnston BC, Thorlund K, Schünemann HJ, et al. Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units. *Health Qual Life Outcomes* 2010;8:116. doi:10.1186/1477-7525-8-116
18 Johnston BC, Thorlund K, da Costa BR, Furukawa TA, Guyatt GH. New methods can extend the use of minimal important difference units in meta-analyses of continuous outcome measures. *J Clin Epidemiol* 2012;65:817-26. doi:10.1016/j.jclinepi.2012.02.008
19 Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. *J Clin Epidemiol* 1994;47:81-7. doi:10.1016/0895-4356(94)90036-1
20 Juniper EF, Guyatt GH, Griffith LE, Ferrie PJ. Interpretation of rhinoconjunctivitis quality of life questionnaire data. *J Allergy Clin Immunol* 1996;98:843-5. doi:10.1016/S0091-6749(96)70135-5
21 Lacasse Y, Wong E, Guyatt G. A systematic overview of the measurement properties of the Chronic Respiratory Questionnaire. *Can Respir J* 1997;4. doi:10.1155/1997/717139
22 Norman GR, Sridhar FG, Guyatt GH, Walter SD. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. *Med Care* 2001;39:1039-47. doi:10.1097/00005650-200110000-00002
23 Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. *J Clin Epidemiol* 1996;49:1215-9. doi:10.1016/S0895-4356(96)00206-5
24 Redelmeier DA, Guyatt GH, Goldstein RS. On the debate over methods for estimating the clinically important difference. *J Clin Epidemiol* 1996;49:1223-4. doi:10.1016/S0895-4356(96)00208-9
25 Sloan JA, Cella D, Frost M, Guyatt GH, Sprangers M, Symonds T, Clinical Significance Consensus Meeting Group. Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. *Mayo Clin Proc* 2002;77:367-70. doi:10.4065/77.4.367
26 Sloan JA, Frost MH, Berzon R, et al, Clinical Significance Consensus Meeting Group. The clinical significance of quality of life assessments in oncology: a summary for clinicians. *Support Care Cancer* 2006;14:988-98. doi:10.1007/s00520-006-0085-y
27 Turner D, Schünemann HJ, Griffith LE, et al. Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. *J Clin Epidemiol* 2009;62:374-9. doi:10.1016/j.jclinepi.2008.07.009
28 Turner D, Schünemann HJ, Griffith LE, et al. The minimal detectable change cannot reliably replace the minimal important difference. *J Clin Epidemiol* 2010;63:28-36. doi:10.1016/j.jclinepi.2009.01.024
29 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:e015587. doi:10.1136/bmjopen-2016-015587
30 De Vet HC, Terwee CB, Mokkink LB, Knol DL. Concepts, theories and models, and types of measurements. In: *Measurement in medicine: a practical guide.* Cambridge University Press, 2011:7-29. doi:10.1017/CBO9780511996214.003
31 Shoukri MM. *Measures of interobserver agreement and reliability.* CRC press, 2010. doi:10.1201/b10433
32 Heppner PP, Wampold BE, Kivlighan DJCL. Research design in counseling: research, statistics, & program evaluation. 2007
33 Kaplan RM, Saccuzzo DP. *Psychological testing: principles, applications, and issues.* Nelson Education, 2017.
34 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420-8. doi:10.1037/0033-2909.86.2.420
35 Arpinelli F, Bamfi F. The FDA guidance for industry on PROs: the point of view of a pharmaceutical company. *Health Qual Life Outcomes* 2006;4:85. doi:10.1186/1477-7525-4-85
36 Barrett B, Brown D, Mundt M, Brown R. Sufficiently important difference: expanding the framework of clinical significance. *Med Decis Making* 2005;25:250-61. doi:10.1177/0272989X05276863
37 Beaton DE, Bombardier C, Katz JN, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. *J Rheumatol* 2001;28:400-5.
38 Beninato M, Portney LG. Applying concepts of responsiveness to patient management in neurologic physical therapy. *J Neurol Phys Ther* 2011;35:75-81. doi:10.1097/NPT.0b013e318219308c
39 Chuang-Stein C, Kirby S, Hirsch I, Atkinson G. The role of the minimum clinically important difference and its impact on designing a trial. *Pharm Stat* 2011;10:250-6. doi:10.1002/pst.459
40 Copay AG, Subach BR, Glassman SD, Polly DWJr, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J* 2007;7:541-6. doi:10.1016/j.spinee.2007.01.008
41 Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003;56:395-407. doi:10.1016/S0895-4356(03)00044-1

42 de Vet HC, Ostelo RW, Terwee CB, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res* 2007;16:131-42. doi:10.1007/s11136-006-9109-9
43 de Vet HC, Terluin B, Knol DL, et al. Three ways to quantify uncertainty in individually applied "minimally important change" values. *J Clin Epidemiol* 2010;63:37-45. doi:10.1016/j.jclinepi.2009.03.011
44 Farivar SS, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in HRQOL scores?*Expert Rev Pharmacoecon Outcomes Res* 2004;4:515-23. doi:10.1586/14737167.4.5.515
45 Gatchel RJ, Mayer TG, Chou R. What does/should the minimum clinically important difference measure? A reconsideration of its clinical value in evaluating efficacy of lumbar fusion surgery. *Clin J Pain* 2012;28:387-97. doi:10.1097/AJP.0b013e3182327f20
46 Guyatt G, Schunemann H. How can quality of life researchers make their work more useful to health workers and their patients?*Qual Life Res* 2007;16:1097-105. doi:10.1007/s11136-007-9223-3
47 Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD* 2005;2:63-7. doi:10.1081/COPD-200050663
48 Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it?*Pharmacoeconomics* 2000;18:419-23. doi:10.2165/00019053-200018050-00001
49 Kamper SJ, Ostelo RW, Knol DL, Maher CG, de Vet HC, Hancock MJ. Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *J Clin Epidemiol* 2010;63:760-766.e1. doi:10.1016/j.jclinepi.2009.09.009
50 Kemmler G, Giesinger J, Holzner BJJ. Clinically relevant, statistically significant, or both? Minimal important change in the individual subject revisited. *J Clin Epidemiol* 2011;64:1467-68.
51 King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res* 2011;11:171-84. doi:10.1586/erp.11.9
52 Kirby S, Chuang-Stein C, Morris M. Determining a minimum clinically important difference between treatments for a patient-reported outcome. *J Biopharm Stat* 2010;20:1043-54. doi:10.1080/10543400903315757
53 Koynova D, Lühmann R, Fischer R. A framework for managing the minimal clinically important difference in clinical trials. *Ther Innov Regul Sci* 2013;47:447-54. doi:10.1177/2168479013487541
54 Leidy NK, Wyrwich KW. Bridging the gap: using triangulation methodology to estimate minimal clinically important differences (MCIDs). *COPD* 2005;2:157-65. doi:10.1081/COPD-200050508
55 Lemieux J, Beaton DE, Hogg-Johnson S, Bordeleau LJ, Goodwin PJ. Three methods for minimally important difference: no relationship was found with the net proportion of patients improving. *J Clin Epidemiol* 2007;60:448-55. doi:10.1016/j.jclinepi.2006.08.006
56 Molnar FJ, Man-Son-Hing M, Fergusson D. Systematic review of measures of clinical significance employed in randomized controlled trials of drugs for dementia. *J Am Geriatr Soc* 2009;57:536-46. doi:10.1111/j.1532-5415.2008.02122.x
57 Osoba D, Bezjak A, Brundage M, Zee B, Tu D, Pater J. Analysis and interpretation of health-related quality-of-life data from clinical trials: basic approach of The National Cancer Institute of Canada Clinical Trials Group. *Eur J Cancer* 2005;41:280-87.
58 Rennard SI. Minimal clinically important difference, clinical perspective: an opinion. *COPD* 2005;2:51-5. doi:10.1081/COPD-200050641
59 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61:102-9. doi:10.1016/j.jclinepi.2007.03.012
60 Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR, Aaronson NK. Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes* 2006;4:70. doi:10.1186/1477-7525-4-70
61 Sloan JA. Assessing the minimally clinically significant difference: scientific considerations, challenges and solutions. *COPD* 2005;2:57-62. doi:10.1081/COPD-200053374
62 Terwee CB, Roorda LD, Dekker J, et al. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol* 2010;63:524-34. doi:10.1016/j.jclinepi.2009.08.010
63 Tubach F, Giraudeau B, Ravaud P. The variability in minimal clinically important difference and patient acceptable symptomatic state values did not have an impact on treatment effect estimates. *J Clin Epidemiol* 2009;62:725-8. doi:10.1016/j.jclinepi.2008.09.012
64 Wright A, Hannon J, Hegedus EJ, Kavchak AE. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). *J Man Manip Ther* 2012;20:160-6. doi:10.1179/2042618612Y.0000000001
65 Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T, Clinical Significance Consensus Meeting Group. Estimating clinically significant differences in quality of life outcomes. *Qual Life Res* 2005;14:285-95. doi:10.1007/s11136-004-0705-2
66 Yost KJ, Eton DT. Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. *Eval Health Prof* 2005;28:172-91. doi:10.1177/0163278705275340
67 Zannikos S, Lee L, Smith HE. Minimum clinically important difference and substantial clinical benefit: Does one size fit all diagnoses and patients?*Semin Spine Surg* 2014;26:8-11. doi:10.1053/j.semss.2013.07.004
68 Schünemann HJ, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. Measurement properties and interpretability of the Chronic Respiratory Disease Questionnaire (CRQ). *COPD* 2005;2:81-9. doi:10.1081/COPD-200050651
69 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:e20160833. doi:10.1542/peds.2016-0833
70 Deyo RA, Inui TS. Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Serv Res* 1984;19:275-89.
71 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10:407-15. doi:10.1016/0197-2456(89)90005-6
72 Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. *J Clin Epidemiol* 1995;48:1369-78. doi:10.1016/0895-4356(95)00054-2
73 Thompson MS, Read JL, Hutchings HC, Paterson M, Harris EDJr. The cost effectiveness of auranofin: results of a randomized clinical trial. *J Rheumatol* 1988;15:35-42.
74 Ware JJr, Keller S. *Interpreting general health measures. Quality of life and pharmacoeonomics in clinical trials.* Lippincott-Raven Publishers, 1995.
75 King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual Life Res* 1996;5:555-67. doi:10.1007/BF00439229
76 Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* 1981;19:787-805. doi:10.1097/00005650-198108000-00001
77 Idler EL, Angel RJ. Self-rated health and mortality in the NHANES-I Epidemiologic Follow-up Study. *Am J Public Health* 1990;80:446-52. doi:10.2105/AJPH.80.4.446
78 Mossey JM, Shapiro E. Self-rated health: a predictor of mortality among the elderly. *Am J Public Health* 1982;72:800-8. doi:10.2105/AJPH.72.8.800
79 Ware JEJr, Manning WGJr, Duan N, Wells KB, Newhouse JP. Health status and the use of outpatient mental health services. *Am Psychol* 1984;39:1090-100. doi:10.1037/0003-066X.39.10.1090
80 Brook RH, Ware JEJr, Rogers WH, et al. Does free care improve adults' health? Results from a randomized controlled trial. *N Engl J Med* 1983;309:1426-34. doi:10.1056/NEJM198312083092305
81 Fayers PM, Machin D. *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes.* John Wiley & Sons, 2013.
82 Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. *Qual Life Res* 2002;11:207-21. doi:10.1023/A:1015276414526
83 Guyatt GH, Jaeschke RJ. Reassessing quality-of-life instruments in the evaluation of new drugs. *Pharmacoeconomics* 1997;12:621-6. doi:10.2165/00019053-199712060-00002
84 Colwell HH, Mathias SD, Turner MP, et al. Psychometric evaluation of the FACT Colorectal Cancer Symptom Index (FCSI-9): reliability, validity, responsiveness, and clinical meaningfulness. *Oncologist* 2010;15:308-16. doi:10.1634/theoncologist.2009-0034
85 Hägg O, Fritzell P, Nordwall A, Swedish Lumbar Spine Study Group. The clinical importance of changes in outcome scores after treatment for chronic low back pain. *Eur Spine J* 2003;12:12-20. doi:10.1007/s00586-002-0464-0
86 Hayes RP, Schultz EM, Naegeli AN, Curtis BH. Test-retest, responsiveness, and minimal important change of the ability to perform physical activities of daily living questionnaire in individuals with type 2 diabetes and obesity. *Diabetes Technol Ther* 2012;14:1118-25. doi:10.1089/dia.2012.0123
87 Schmitt J, Di Fabio RP. The validity of prospective and retrospective global change criterion measures. *Arch Phys Med Rehabil* 2005;86:2270-6. doi:10.1016/j.apmr.2005.07.290
88 Akl EA, Sun X, Busse JW, et al. Specific instructions for estimating unclearly reported blinding status in randomized trials were

reliable and valid. *J Clin Epidemiol* 2012;65:262-7. doi:10.1016/j.jclinepi.2011.04.015
89 Furukawa TA, Jaeschke R, Cook D, Guyatt G. Measuring of patients' experience. In: *Users' guides to the medical literature: a manual for evidence-based clinical practice.* McGraw-Hill, 2008: 249-72.
90 Levine M, Ioannidis J, Haines T, Guyatt G. Harm (observational studies). In: *Users' guides to the medical literature: a manual for evidence-based clinical practice.* McGraw-Hill, 2008.
91 Randolph A, Cook D, Guyatt G. Prognosis. In: *Users' guides to the medical literature: a manual for evidence-based clinical practice.* McGraw-Hill, 2008.
92 Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010;340:c117. doi:10.1136/bmj.c117
93 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:e028777.
94 Burris HA3rd, Lebrun F, Rugo HS, et al. Health-related quality of life of patients with advanced breast cancer treated with everolimus plus exemestane versus placebo plus exemestane in the phase 3, randomized, controlled, BOLERO-2 trial [correction: *Cancer* 2019;125:1387-88]. *Cancer* 2013;119:1908-15. doi:10.1002/cncr.28010
95 Nonoyama ML, Brooks D, Guyatt GH, Goldstein RS. Effect of oxygen on health quality of life in patients with chronic obstructive pulmonary disease with transient exertional hypoxemia. *Am J Respir Crit Care Med* 2007;176:343-9. doi:10.1164/rccm.200702-308OC
96 Mills KA, Naylor JM, Eyles JP, Roos EM, Hunter DJ. Examining the minimal important difference of patient-reported outcome measures for individuals with knee osteoarthritis: a model using the knee injury and osteoarthritis outcome score. *J Rheumatol* 2016;43:395-404. doi:10.3899/jrheum.150398
97 Olsen MF, Bjerre E, Hansen MD, Tendal B, Hilden J, Hróbjartsson A. Minimum clinically important differences in chronic pain vary considerably by baseline pain and methodological factors: systematic review of empirical studies. *J Clin Epidemiol* 2018;101:87-106.e2. doi:10.1016/j.jclinepi.2018.05.007
98 Terluin B, Eekhout I, Terwee CB. The anchor-based minimal important change, based on receiver operating characteristic analysis or predictive modeling, may need to be adjusted for the proportion of improved patients. *J Clin Epidemiol* 2017;83:90-100. doi:10.1016/j.jclinepi.2016.12.015
99 Terluin B, Eekhout I, Terwee CB, de Vet HC. Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. *J Clin Epidemiol* 2015;68:1388-96. doi:10.1016/j.jclinepi.2015.03.015

Web appendix 1: Search strategy for Medline, January 1989 to April 2015
Web appendix 2: Credibility instrument for judging the trustworthiness of minimal important difference (MID) estimates
Web appendix 3: Application of the Credibility instrument for judging the trustworthiness of minimal important difference estimates–worked examples
Web appendix 4: Data extraction sheet with sample extraction