THE SELECTION OF OPTIMAL ANCHOR-BASED MINIMAL IMPORTANCE DIFFERENCE

THE SELECTION OF OPTIMAL ANCHOR-BASED MINIMAL IMPORTANCE DIFFERENCE TO ENHANCE TRUSTWORTHY INTERPRETATION OF PATIENT-REPORTED OUTCOMES IN CLINICAL RESEARCH AND EVIDENCE-BASED DECISION-MAKING

By Yuting Wang, MD

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy

McMaster University © Copyright by Yuting Wang, 2024

McMaster University DOCTOR OF PHILOSOPHY (2024)

Hamilton, Ontario (Health Research Methodology)

TITLE: The Selection of Optimal Anchor-based Minimal Important Difference to Enhance Trustworthy Interpretation of Patient-reported Outcomes in Clinical Research and Evidence-based Decision Making

AUTHOR: Yuting Wang, MD

SUPERVISOR: Dr. Gordon H Guyatt, Distinguished Professor

NUMBER OF PAGES: xiv, 131

ABSTRACT

Patient centered research encourages incorporating patient perspectives to inform the pursuit of clinical questions, the conduct of clinical research, the benefit-risk assessments and decision-making, as well as the delivery of appropriate health care. Researchers have thus increasingly adopted Patient-reported outcomes measurements (PROMs)—the instruments that capture patient-reported outcomes (PROs)—to gather the important outcomes from patients' viewpoint without outside interpretation from anyone else, such as quality of life, mental health, and physical function. Typically, researchers use pre- and post-event PROM results to measure the impacts of an intervention, informing benefit-risk assessments and decision-making. The interpretation of the PROM results involves deciding whether a particular treatment effect is trivial, small but important, moderate or large. To aid such interpretation, researchers proposed anchor-based minimal important difference (MID): the smallest important difference or change, either beneficial or harmful, that patients perceive as important. The trustworthy interpretation, however, relies largely on the choice of an optimal anchor-based MID. With the widespread recognition of the usefulness of anchor-based MIDs, the number of published anchor-based MIDs for PROMs has grown rapidly. Though all the published MIDs have gone through peer review, considerable difficulties for optimal MID selection exist because the MIDs for a given PROM vary widely and are not equally trustworthy in one way or another. This thesis aims to address the use of anchor-based MIDs in enhancing the interpretation of PROMs, with particular focuses on the methodological issues related to selecting an optimal MID and the development of a systematic approach to selecting an optimal MID. To start with, we conducted a systematic survey of the literature addressing the

issues related to selecting optimal MIDs. Subsequently, based on the survey information, we refined the existing anchor-based MID credibility instrument by adding the construct proximity assessment, which supplemented the MID methodology assessments. Then, informed by the work above, we developed the systematic approach to selecting an optimal anchor-based MID for a given PROM, which is geared by both the methodology rigor and application contexts. Finally, this thesis concludes with insights to the application of the selection approach and the opportunities for future research.

ACKNOWLEGEMENTS

First and foremost, I would like to thank Dr. Gordon Guyatt for his outstanding mentorship. Gordon, you have provided me so many cherished opportunities that opened my academic world. Without your support and guidance, I could not have accomplished all that I have thus far in my academic career. I am so impressed by your passion towards teaching and life-long learning. You make me believe that someone can be born for academics and human being's brain may work better by eating and sleeping less.

To my supervisory committee, Drs. Michael Walsh and Jason Busse, you have been invaluable sources of knowledge, expertise and inspiration. Thank you for the insightful feedback on this work and for providing me with advice that I will take with me throughout my career.

I feel honored to complete my PhD training in the Health Research Methodology (HRM) program, the Department of Health Research Methods, Evidence and Impact (HEI), at the birthplace of evidence-based medicine among world-renowned experts. The program has exceeded all my expectations and is truly an exceptional institution to complete graduate training.

To all the co-authors that I have had the privilege of collaborating with, your contributions have been instrumental to this work, and I am truly appreciative of your commitments and efforts. In particular, Drs. Tahira Devji and Alonso Carrasco-Labra, you both have provided important guidance and feedback to improve the design and conduct of the projects. The members of the

project steering committee, Drs. Madeleine King; Berend Terluin; Caroline Terwee and Toshi Furukawa, thanks for your expertise and I really appreciate your flexibility: it was difficult to schedule meetings across the four continents.

To my parents, Fusheng Wang and Jianying He, thank you for your constant love and encouragement. You have never hesitated for a second about supporting my decision of pursuing PhD abroad. You have relieved me of all the external pressure to help me achieve my goals. Without your understanding, I will not be able to arrive the destination of this long journey—my 14 years post-secondary education. I love you both.

To my dear husband, Xiang, thanks for bearing my tempers and always being supportive. I don't think there is anyone else who can be such a supportive husband like you. My graduation will meet our 8th anniversary, which is the best number in China. I am so willing to spend the rest of my life with you. To my sweety, Nemy, thanks for accompanying me in Canada. Your sweet laugh really warms me during the long, cold winter. Mama hopes you will be always happy.

To my friend, Qiukui, you led me to the field of evidence-based medicine. To me, you are both my mentor and friend. You are generous to give a helping hand and have helped me through several difficulties. Hope you everything goes well in Canada.

To our friends, Emma, Xiaoqin, Liang and Xue, thank you to make our life (Nemy and me) in Canada easier and full of happiness. Nemy loves you all and I am grateful for your love and care to Nemy.

Last but not least, to McMaster Student Union Child Center, without your day care for Nemy, I can imagine how messy my life in Canada will be and it will be impossible for me to manage both my PhD study and care for Nemy. Thanks very much, Ms. Michelle, Marley, and Alisia. You have left very wonderful memories to Nemy. We will miss you in China.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

Abbreviation	Definition
BMJ	British Medical Journal
CI	Confidence Interval
JCE	Journal of Clinical Epidemiology
KOOS	Knee Injury and Osteoarthritis Outcome Score
MID	Minimal Important Difference
VAS	Visual Analogue Scale
PRO	Patient Reported Outcome
PROM	Patient Reported Outcome Measurement
PROMID	Patient Reported Outcome Minimal Important Difference Database
SF-36-MCS	36-item Short Form Survey-Mental Component Summary
WOMAC	The Western Ontario and McMaster Universities Osteoarthritis Index

DECLARATION OF ACADEMIC ACHIEVEMENT

Chapter 1. This chapter is unpublished. YW is the sole author.

Chapter 2: This chapter is published in JCE. YW, GHG and TD conceived the study idea. YW, ACL, BT, CT, GHG, MK, MW, TD, and TF developed the protocol and had oversight of the review process. YW, AN, LECL, MB, MG, MP, MRP, QH, TD, YX, YW and VW conducted screening for eligible articles. YW, TD created the data abstraction form and the initial code list for data abstraction. YW, AQ, MB, MP, QH, and VW extracted data and conducted coding. YW, GHG developed items. YW wrote the first draft of the manuscript. YW, ACL, BT, CT, GHG, MK, MW, TD and TF interpreted the data analysis, critically commented on the manuscript drafts. YW, ACL, AN, AQ, BT, CT, GHG, LECL, MB, MK, MG, MP, MRP, MW, QH, TD, TF, YX, YW and VW reviewed, revised, and approved the manuscript.

Chapter 3: This chapter is published in JCE. YW, GHG, TD, TAF and YT conceived the study idea. YW, ACL, AQ, GHG and TD conducted series of discussions for the development of the item and judgement principles. BT, CT, MK, MW, TAF, YT reviewed, added suggestions to and ultimately approved the final version of the item wording, response options and judgement principles. YW and TD conducted the introduction sessions for raters. YW, AQ, EK, QH, ND and TD assessed the samples. YW wrote the first draft of the manuscript. ACL, AQ, BT, CT, EK, GHG, QH, MK, MW, TD, TAF and YT critically commented on the manuscript drafts. YW,

ACL, AQ, BT, CT, EK, GHG, QH, MK, MW, TD, TAF and YT reviewed, revised, and approved the manuscript.

Chapter 4: This chapter is published in the BMJ. YW, GHG and TD initiated the project. YW, AC-L, BT, CBT, GHG, MTK, MW, TD and TAF provided insights on the selection framework and contributed to the consensus of the items used to develop the selection approach. YW, AC-L, TD and GHG drafted the details of the selection approach. All authors provided feedback for the revision of the selection approach. YW was responsible for revising the selection approach until all authors reached agreements. YW wrote the initial draft of the manuscript and all other authors reviewed and revised the manuscript draft. All authors approved the final version of the manuscript.

Chapter 5: This chapter is unpublished. YW is the sole author.

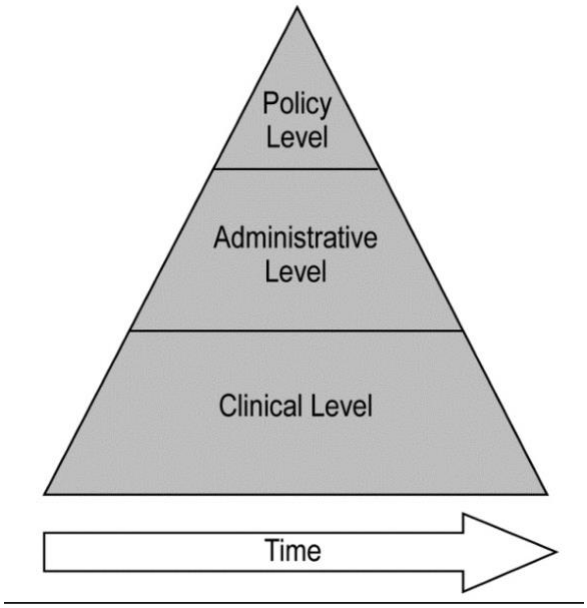
Chapter 1: Introduction of the Thesis

Current health research has been shifting the focus from the perspectives of researchers or clinicians to the perspectives of patients, referring to as patient-centered research [1-4]. Patient-centered research believes that patients have unique perspectives that can change the pursuit of clinical questions and thus improve clinical research [4]. Patients' perspectives provide pivotal information of the patients' unmet needs and the treatments effects known only to patient for heath researchers, health care professionals and policy makers to refine health research, form high quality treatment guidelines and finally deliver appropriate care [1].

While many authorities have increasingly recognized the importance of patient perspectives and advocated the use of patient perspectives [5-8], incorporating patient perspectives in health research remains challenging. Patient-reported outcomes (PROs), measured by patient-reported outcomes measurements (PROMs) via a set of items, directly capture the information on the important outcomes, including symptom status, physical function, mental health, social function and wellbeing, from the patients' and caregivers' viewpoint without outside interpretation from anyone else [5]. Using PROMs thus constitutes an effective strategy to incorporate patient perspectives [9-11].

We can collect PROM information at various levels for a range of different purposes, from clinical to policy making (**Figure 1**) [11 12]. In some cases, using pre- and post-event PROMs can help measure the impacts of an intervention. The systematic use of such PROM information in daily clinical practice contributes to better communication and decision making between clinicians and patients and improves quality of care [10 13-15]. In clinical research, the results of PROMs lead clinical guideline developers, authorizations of medicines and health system policymakers to make more informed benefit-risk assessments for the candidate interventions [16-18].

Figure 1. PROM Information Pyramid



Source
Canadian Institute of Health Information [11 12]

The trustworthy interpretation of PORM results is, however, challenging. Typically, PROMs, either generic (applied across different populations) or condition-specific (used to assess outcomes that are specific or unique to particular diseases or sectors of care), assess a patient's health status at a particular point in time, of which the change scores between two time points reflect improvement or deterioration of the patient's health status. Though we can easily obtain the change scores by simply administrating the PROMs twice, how much importance patients attach to these changes—trivial, small but important, moderate or large changes—proves difficult to interpret.

To aid such interpretation, researchers proposed the minimal important difference (MID): the smallest important difference or change, either beneficial or harmful, that patients perceive as important [19 20]. When estimating MIDs, investigators commonly use two approaches: anchor-based

and distribution-based methods. Anchor-based MIDs relate a difference in the target PROM to an independent measure (i.e., the anchor) that is itself interpretable [19]. Distribution-based methods rely on statistical characteristics of PROM scores in a study sample, providing no clear relation to the importance of the change in PROM scores to patients [21]. Anchor-based MIDs therefore represent a far better approach to aid the interpretation [21 22].

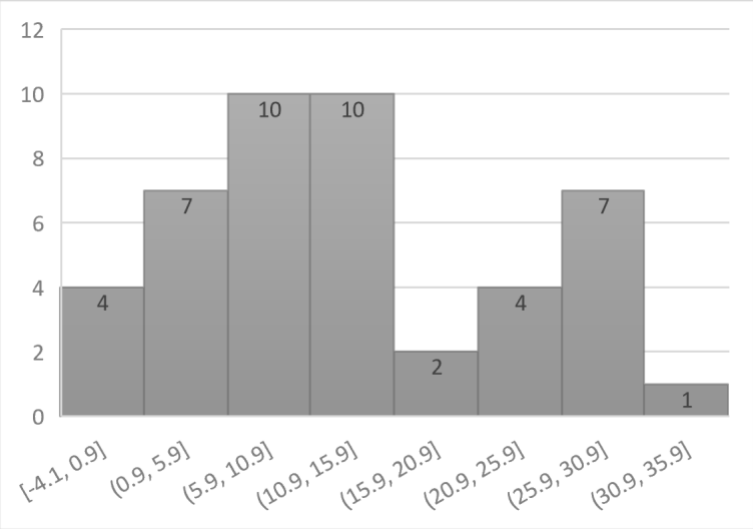
With the widespread recognition of the usefulness of anchor-based MIDs, the number of published studies providing anchor-based MIDs for PROMs has grown rapidly [23]. As a result, for a given PROM, multiple anchor-based MIDs exist [24 25]. As we have discovered in developing a comprehensive inventory of anchor-based MIDs for PROMs [26], however, those MID estimates differ substantially (see the example in **Figure 2**), which in turn, adds difficulties to the trustworthy interpretion of PROM results.

Previously, we developed the anchor-based MID credibility assessment instrument [27], aiming to differentiate trustworthy from untrustworthy anchor-based MIDs. The development of the credibility instrument did not, however, solve the problem of choosing an optimal MID estimate for the interpretation of the PROM results: when a number of widely varying MID estimates are available, credibility may be similar across the estimates (see the example in **Figure 3**).

Therefore, to enhance the interpretation of PROM results and better understand patients' perspectives in clinical research and evidence-based decision-making, a need still exists for a guidance on how to identify, from among the many, the optimal MID. This thesis aims to address the use of anchor-based MIDs, with particular focuses on the methodological issues related to

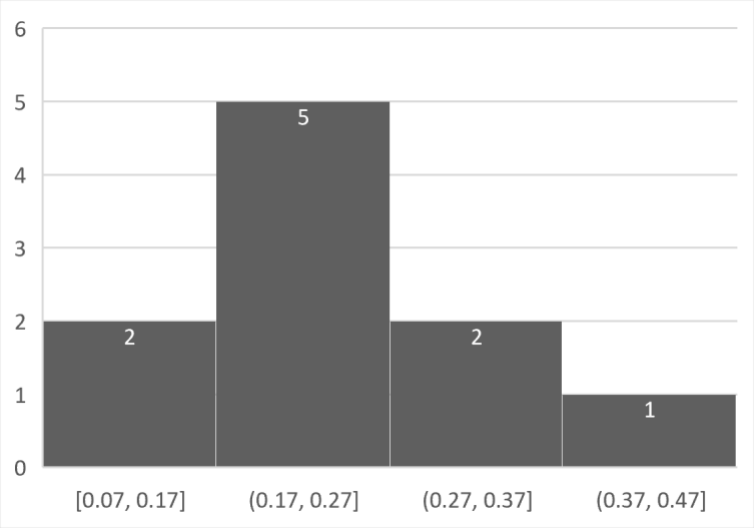
selecting an optimal MID and the development of a systematic approach to selecting an optimal MID.

Figure 2. Wide variation among all available anchor-based MIDs for WOMAC-pain (up to 2018)



Note: Data were drawn from the anchor-based MID inventory [26]. A total of 45 absolute MIDs. WOMAC= The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC).

Figure 3. Wide variation among all available anchor-based MIDs with the highest credibility ratings for EQ-5D (up to 2018)



Note: Data were drawn from the anchor-based MID inventory [26]. A total of 10 absolute MIDs with the highest credibility ratings.
EQ-5D= EuroQol-5D Utility Index.

Chapter 2 documents a systematic survey of the literature addressing the reporting of studies estimating anchor-based MIDs and the choice of optimal anchor-based MIDs. This systematic survey qualitatively summarized the literature and identified the items related to selecting optimal MIDs. These items provide a conceptual framework to inform the following refinement of the MID credibility assessment as well as the development of a selection approach for optimal MIDs.

Chapter 3 informed by the information from Chapter 2, reports the refinement of the existing anchor-based MID credibility assessment instrument [27]. We extended the anchor-based MID credibility instrument by adding an item addressing construct proximity as an alternative to the correlation item—the correlation between the PROM and the anchor—to resolve the deficiency of the assessments of the correlation item due to the underreporting in most MID studies.

Chapter 4, informed by information from Chapter 2 and 3, describes the methods of the development of the systematic approach to selecting an optimal anchor-based MID for a given PROM, the rationale for the approach, and the detailed steps to selecting the optimal MID from available MID estimates. The approach is geared to explaining the variability of the MIDs for the PROM of interest by the methodological rigor and contextualized factors influencing the MID application, and where appropriate, provides a single optimal MID, i.e., the median of the selected estimates in a relatively narrow range.

This thesis ends with **Chapter 5**, which is a discussion of the previous chapters, summarizing the main findings, listing the strengths, and addressing the limitations, providing implications for practice and an exploration of opportunities and directions for future research.

Reference

1. de Wit M, Cooper C, Reginster JY. Practical guidance for patient-centred health research. *Lancet* 2019;393(10176):1095-96.
2. Canadian Institutes of Health Research. Canada's Strategy for Patient-Oriented Research: improving health Outcomes through evidence-informed care. August 2011. [Accessed April 28, 2023]. Available from: https://cihr-irsc.gc.ca/e/44000.html#f1
3. Bardes CL. Defining "patient-centered medicine". *N Engl J Med* 2012;366(9):782-3.
4. Frank L, Basch E, Selby JV. The PCORI perspective on patient-centered outcomes research. *Jama* 2014;312(15):1513-4.
5. Food and Drug Administration. Guidance for Industry: patient-reported outcome measures: use in medical product development to support labelling claims. 2009. [Accessed June 28, 2022]. Available from: https://www.fda.gov/media/77832/download
6. Coens C, Pe M, Dueck AC, Sloan J, Basch E, Calvert M, Campbell A, Cleeland C, Cocks K, Collette L, Devlin N, Dorme L, Flechtner HH, Gotay C, Griebsch I, Groenvold M, King M, Kluetz PG, Koller M, Malone DC, Martinelli F, Mitchell SA, Musoro JZ, O'Connor D, Oliver K, Piault-Louis E, Piccart M, Quinten C, Reijneveld JC, Schürmann C, Smith AW, Soltys KM, Taphoorn MJB, Velikova G, Bottomley A. International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium. *Lancet Oncol* 2020;21(2):e83-e96.
7. European Medicines Agency. Patient experience data in EU medicines development and regulatory decision-making. Accessed November 7, 2022. https://www.ema.europa.eu/en/documents/other/executive-summary-patient-experience-data-eu-medicines-development-regulatory-decision-making_en.pdf5.
8. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *Jama* 2013;309(8):814-22.
9. Terner M, Louie K, Chow C, Webster G. Advancing PROMs for health system use in Canada and beyond. *J Patient Rep Outcomes* 2021;5(Suppl 2):94.
10. Nelson EC, Eftimovska E, Lind C, Hager A, Wasson JH, Lindblad S. Patient reported outcome measures in practice. *Bmj* 2015;350
11. Canadian Institutes for Health Information. PROMs background document. 2015. [Accessed April 28.2023].Available from: https://www.cihi.ca/sites/default/files/proms_background_may21_en-web_0.pdf.
12. Canadian Institutes for Health Information. Health Outcomes of Care: An idea whose time has come. [Accessed April 28.2023].Available from: https://secure.cihi.ca/free_products/HealthOutcomes2012_EN.pdf
13. Valderas JM, Kotzeva A, Espallargues M, Guyatt G, Ferrans CE, Halyard MY, Revicki DA, Symonds T, Parada A, Alonso J. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual Life Res* 2008;17(2):179-93.
14. Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: a structured review. *Journal of evaluation in clinical practice* 2006;12(5):559-68.
15. Chen J, Ou L, Hollis SJ. A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. *BMC health services research* 2013;13(1):1-24.
16. Black N. Patient reported outcome measures could help transform healthcare. *Bmj* 2013;346:f167.
17. Kluzek S, Dean B, Wartolowska KA. Patient-reported outcome measures (PROMs) as proof of treatment efficacy. *BMJ Evid Based Med* 2022;27(3):153-55.
18. Food and Drug Administration. Guidance for Industry: Benefit-Risk Assessment for New Drug and Biological Products Guidance for Industry (Draft Guidance). 2021. [Accessed June 28, 2022]. Available from: chrome-

extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.fda.gov/media/152544/download
19. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10(4):407-15.
20. Schünemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res* 2005;40(2):593-7.
21. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, Guyatt GH. The minimal detectable change cannot reliably replace the minimal important difference. *J Clin Epidemiol* 2010;63(1):28-36.
22. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *Jama* 2014;312(13):1342-3.
23. Devji T, Carrasco-Labra A, Guyatt G. Mind the methods of determining minimal important differences: three critical issues to consider. *BMJ Ment Health* 2021;24(2):77-81.
24. Devji T, Guyatt GH, Lytvyn L, Brignardello-Petersen R, Foroutan F, Sadeghirad B, Buchbinder R, Poolman RW, Harris IA, Carrasco-Labra A, Siemieniuk RAC, Vandvik PO. 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.
25. Hao Q, Devji T, Zeraatkar D, Wang Y, Qasim A, Siemieniuk RAC, Vandvik PO, Lähdeoja T, Carrasco-Labra A, Agoritsas T, Guyatt G. 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.
26. Carrasco-Labra A, Devji T, Qasim A, Phillips MR, Wang Y, Johnston BC, Devasenapathy N, Zeraatkar D, Bhatt M, Jin X, Brignardello-Petersen R, Urquhart O, Foroutan F, Schandelmaier S, Pardo-Hernandez H, Hao Q, Wong V, Ye Z, Yao L, Vernooij RWM, Huang H, Zeng L, Rizwan Y, Siemieniuk R, Lytvyn L, Patrick DL, Ebrahim S, Furukawa TA, Nesrallah G, Schünemann HJ, Bhandari M, Thabane L, Guyatt GH. Minimal important difference estimates for patient-reported outcomes: A systematic survey. *J Clin Epidemiol* 2021;133:61-71.
27. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, Zeraatkar D, Bhatt M, Jin X, Brignardello-Petersen R, Urquhart O, Foroutan F, Schandelmaier S, Pardo-Hernandez H, Vernooij RW, Huang H, Rizwan Y, Siemieniuk R, Lytvyn L, Patrick DL, Ebrahim S, Furukawa T, Nesrallah G, Schünemann HJ, Bhandari M, Thabane L, Guyatt GH. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *Bmj* 2020;369:m1714.

Chapter 2: A systematic Survey Identified Methodological Issues in Studies Estimating Anchor-based Minimal Important Differences in Patient-Reported Outcomes

Yuting Wang; Tahira Devji; Anila Qasim; Qiukui Hao; Vanessa Wong; Meha Bhatt; Manya Prasad; Ying Wang, MPharm; Atefeh Noori; Yingqi Xiao; Maryam Ghadimi; Luis Enrique Colunga Lozano; Mark R. Phillips; Alonso Carrasco-Labra; Madeleine King; Berend Terluin; Caroline Terwee; Michael Walsh; Toshi A. Furukawa; Gordon H. Guyatt

Published in: J Clin Epidemiol. 2022 Feb;142:144-151. doi: 10.1016/j.jclinepi.2021.10.028.

ABSTRACT

Objective: To systematically survey the literature addressing the reporting of studies estimating anchor-based minimal important differences (MIDs) and choice of optimal MIDs.

Study design and Setting: We searched Medline, Embase and PsycINFO from 1987 to March 2020. Teams of two reviewers independently identified eligible publications and extracted quotations addressing relevant issues for reporting and/or selecting anchor-based MIDs. Using a coding list, we assigned the same code to quotations capturing similar or related issues. For each code, we generated an 'item', i.e. a specific phrase or sentence capturing the underlying concept. When multiple concepts existed under a single code, the team created multiple items for that code. We clustered codes addressing a broader methodological issue into a 'category' and classified items as relevant for reporting, relevant for selecting an anchor-based MID, or both.

Results: We identified 136 eligible publications that provided 6 categories (MID definition, anchors, patient-reported outcome measures, generalizability, and statistics) and 24 codes. These codes contained 34 items related to reporting MID studies, of which 29 were also related to selecting MIDs.

Conclusion: The systematic survey identified items related to reporting of anchor-based MID studies and selecting optimal MIDs. These provide a conceptual framework to inform the design of studies related to MIDs, and a basis for developing a reporting standard and a selection approach for MIDs.

Keywords: Patient-reported outcome measure, minimal important difference.

What is new?

Key findings

We identified 34 items related to reporting studies estimating anchor-based minimal important differences (MIDs), of which 29 were also related to selecting MIDs.

What this adds to what was known?

In contrast to previous MID reviews, in addition to pointing out methodological issues of anchor-based MIDs, this systematic survey used qualitative synthesis to identify and summarize relevant items related to reporting and selecting anchor-based MIDs.

What is the implication and what should change now?

The systematic survey comprehensively summarizes reported methodological issues in estimating MIDs and thus provides a conceptual framework to inform development of a reporting standard and a systematic selection approach for anchor-based MIDs.

1. Introduction

To address outcomes of importance to patients, clinical trials, systematic reviews, and clinical practice guidelines increasingly rely on patient-reported outcomes (PROs) captured by PRO measures (PROMs). Interpreting PROM results requires understanding whether changes in patients' scores represent trivial, small but important, moderate or large changes. Defining a threshold that represents the smallest important difference or change, either beneficial or harmful, that patients perceive as important - the minimal important difference (MID) – can greatly aid interpretation of study results [1,2].

When estimating MIDs, investigators commonly use two approaches: anchor-based and distribution-based methods [3]. Anchor-based MIDs relate a difference in the target PROM to an independent measure (i.e., the anchor) that is itself interpretable [1]. Distribution-based methods rely on statistical characteristics of PROM scores in a study sample [3]. Because of variability in the relation between anchor-based and statistical methods, and the expectation that statistical approaches will not relate in a predicable way to what patients consider important, anchor-based methods represent the optimal approach to deriving MIDs [4-6].

Increasingly, multiple anchor-based MID estimates exist for popular PROMs. If, however, those estimates differ substantially – and, as we have discovered in developing a comprehensive inventory of anchor-based MIDs for PROMs, they typically do [7]– interpretating PROM results becomes problematic. With an increasing number of published MID estimates, a growing need exists for guidance on how to identify, from among the many, the optimal MID. When ascertaining

the credibility of available MIDs for the inventory [8], we also noted a second problem: inadequate reporting that often made assessments of credibility impossible [9].

Though urgently needed to remedy these problems, a systematic approach to selecting the optimal MIDs from the available estimates, and a reporting standard to which authors of studies estimating anchor-based MIDs could adhere, remain unavailable. With the aim of informing both the selection of optimal MIDs and standards for reporting of MID studies, we conducted a systematic survey of the literature addressing methodological issues in reporting studies estimating anchor-based MIDs and selecting optimal MIDs from a range of options.

2. Methods

The systematic survey consisted of 3 phases: identifying potentially informative articles; data abstraction; and code and item development. A steering committee (ACL, BT, CT, GG, MK, MW, TD, TF and YW) developed the protocol and provided oversight of the review process.

2.1. Eligibility criteria

The steering group formulated the following inclusion criteria:

1. publications developing methods to estimate an anchor-based MID;

2. publications addressing specific methodological issues in estimating an anchor-based MID, such as the validity of anchor questions, statistical calculation methods, interpretation of results, and variability and application of MIDs. For original studies, we required a clear statement of exploring the issues in study objectives;

3. commentaries and critiques focusing on particular anchor-based MIDs and discussing methodological issues;

4. literature reviews, including systematic or narrative reviews, with or without meta-analysis, of established anchor-based MIDs for a PROM including discussions addressing (in)consistency across summarized MIDs or authors' views regarding MID credibility or applicability;

5. publications providing guidance on reporting anchor-based MIDs and/or aiding the choice of an established anchor-based MID among multiple established MIDs.

We excluded conference abstracts, textbooks and non-English language publications.

2.2. Search strategy and article selection

We searched publications in Medline, Embase and PsycINFO from 1987 to March 2020 using the term MID and its variations as key words (*Appendix 1).* Prior to publication selection, reviewers underwent training and calibration exercises in which they screened the same sample of 10 citations, discussed discrepant choices, and developed instructions to minimize subsequent disagreement. Teams of two reviewers independently performed title and abstract screening of citations identified from the literature search and, to determine final eligibility, full-text screening of potentially eligible articles. An arbitrator (YW, TD, or GG) resolved discrepancies during the screening process.

2.3. Data extraction and Code development

Data extraction involved collecting the direct quotations addressing issues relevant for selecting and/or reporting anchor-based MIDs and then assigning a code for each issue. Through detailed review of 21 eligible studies identified from a previous review [8], TD and YW developed draft

codes with detailed instructions for data abstraction. The codes provided a label capturing a collection of similar or related issues.

The steering committee provided feedback and suggestions to reach a first draft of the coding system. After participating in calibration exercises, the reviewers (AQ, MB, MP, QH, YW, VW), working in pairs according to level of experience (i.e., more experienced reviewers (AQ, QH, YW) were paired with less experienced reviewers), identified relevant quotations from eligible publications and matched them to an existing code. When reviewers could not match a quotation to a code, YW and TD determined whether a new code was needed, and if so, added it to a shared online data extraction form accessible to reviewers. This method allowed dynamic updating of the coding system as new data emerged. When the more experienced reviewers considered a question particularly fundamental or challenging, a third investigator (GG, TD) adjudicated. *Appendix 2* presents the workflow of quotation abstraction and code development.

Reviewers also extracted the following information: 1) authors, 2) publication year, 3) country, 4) type of publication: original study, systematic review or methodological article (any eligible article that is neither original study nor systematic review).

2.4. Item Development

To further ensure all concepts had been accurately captured, YW, following data abstraction, checked all codes against the abstracted quotations. Consulting with a senior investigator (GG), YW added new codes to capture any that had been overlooked, and discarded codes that on further reflection provided redundant or insufficiently relevant material. To capture the concept underlying

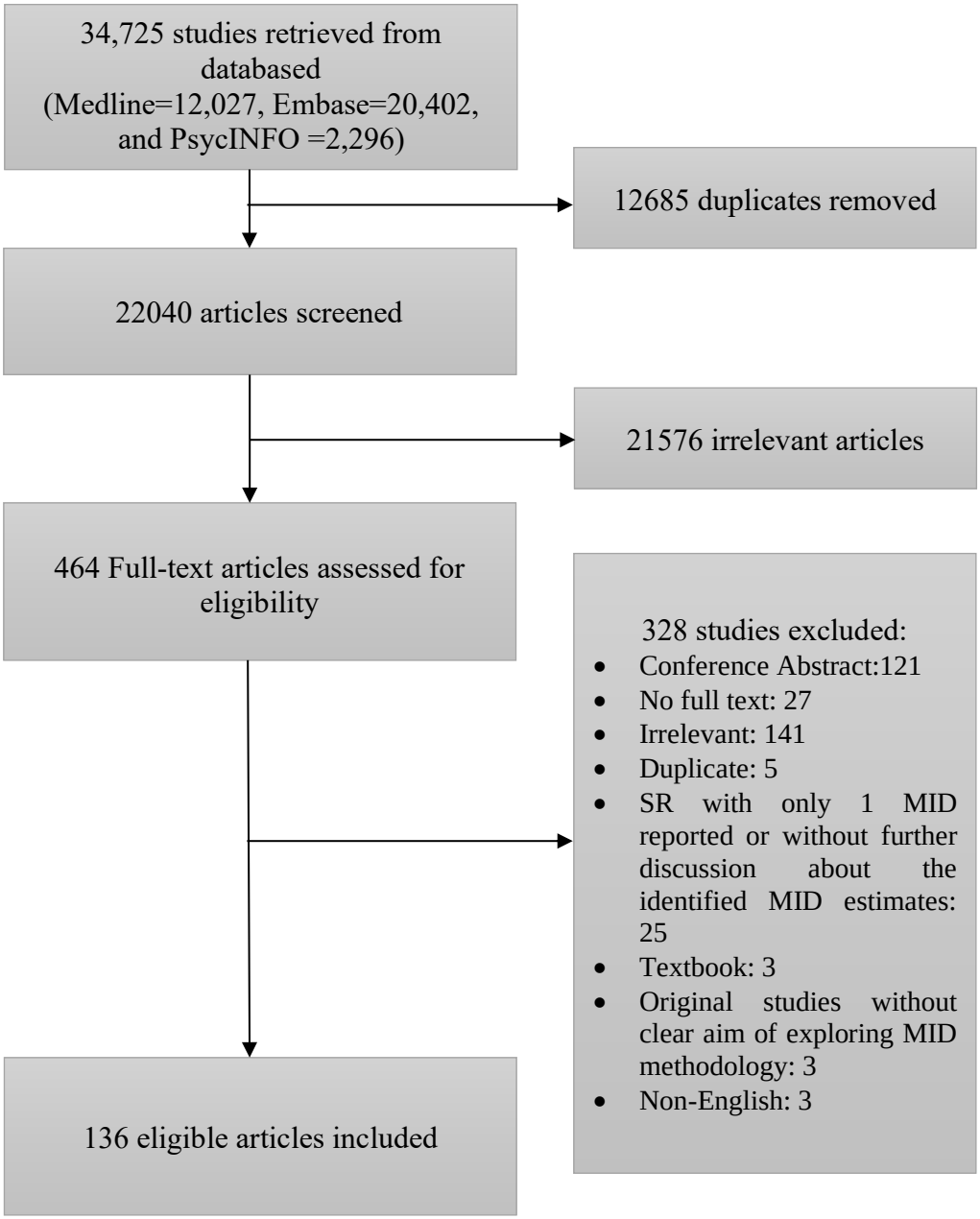
the codes, the team generated, for each code, a more specific phrase or sentence that we called an "item". When multiple concepts existed in the quotations under a single code, the team created more than one items for that code. At least one quotation supported each item. The team then decided which items were relevant for reporting, selecting an anchor-based MID or both. The committee reviewed the penultimate code and item list, and the final codes and items included their suggestions. Following suggestions from the committee, we clustered different codes addressing a same broader methodological issue for anchor-based MID into a "category".

3. Results

3.1. Search results

Of 34,725 records from Medline, Embase and PsycINFO, 22,040 proved, after removing duplicates, unique; of these, reviewers deemed 464 articles eligible for full-text screening, ultimately resulting in 136 eligible publications (*Figure 4*). We included 21 systematic reviews with discussions regarding anchor-based MID methodology and the variability of different MIDs for a given PROM; 40 original studies exploring anchor-based MID methodological issues; and 75 methodological articles including literature reviews, commentaries, critiques and simulation studies *(Table 1*).

Figure 4. Study selection flow chart.



3.2. Items for reporting and/or selecting an anchor-based MID

Most publications appeared after 2010 and provided up to 16 codes *(Table 1*). The team created 61 codes for the initial coding system, adding 2 codes during data abstraction. The codes were

saturated – that is, no new codes emerged - after 30 publications. By checking the initial codes in relation to the abstracted quotation, we discarded 6 codes as irrelevant, and eliminating redundancy led to the reduction in the number of codes from 43 to 10 codes (see *Appendix 3* for details). The refinement of the initial codes thus resulted in 24 final codes. These codes contained 34 items grouped into six categories: definition (2 codes and 2 items); anchor issues (10 codes and 14 items); PROM issues (1 code and 2 items); generalizability of MID (3 codes and 5 items); statistical issues (6 codes and 7 items) and other issues (2 codes and 4 items) (*Table 2*). Definition, anchor issues, generalizability of MIDs, and statistical issues proved the categories most frequently addressed: each was identified in more than 10 articles (Ta*ble 2*). A median of 18 quotations supported the items.

Table 2 presents each category, code, item and *Appendix 4* provides a characteristic example of a quotation for each item. Most items were related to both reporting and selecting anchor-based MID; the team classified 5 items as related only to reporting: MID definition, standardized MID terminology, possible anchors, expressing MID in relative or absolute terms and possible anchor-based MID calculation approaches (*Table 2*).

Table 1. Characteristics of the 136 eligible articles.

Characteristics	Number of articles
Decade of publication	
2010s (up to 2020)	89
2000s	41
1990s	5
1980s	1
Type of publication	
Original studies	40
Systematic reviews	21
Other Methodological articles	75
Number of codes	
1-5	76
6-10	48
11-16	12

Table 2. List of identified Items relevant to reporting and/or selecting anchor-based MID

Category	Code	Item	N ξ
Definition	MID definition	MID definition †	95
	MID terminology	Standardize MID terminology †	6
Anchor issues	Anchor type	Possible anchors†	30
	Anchor perspective	The change being perceived as minimally important on the anchor would be perceived as beneficial from the patient's viewpoint.	30
		Proxy perspective is acceptable if patients cannot make decisions.	4
		Estimating MID should use clinical evaluation as anchor.	19
	Anchor threshold	The choice of what constitutes an MID on the anchor should reflect a small but important difference .	38
	Anchor interpretability	Anchor should be easily understandable.	18
	Correlation between PROM and anchor	At least moderate correlation between PROM and anchor.	36
		Transition scale should correlate equally strongly with the pretest and posttest PROM scores but with opposite signs.	6
		The correlation of transition scale with change scores on PROM should be at least 0.2 greater (in absolute terms) than the correlations with either the baseline or the follow-up test scores.	4

	Anchor construct	Anchor and PROMs should measure the same or closely related constructs.	20
	Recall time of transition scale	The recall time should not be too long when using transition scale as anchor.	53
	Unchanged group of transition scale	The use of unchanged group when using transition scale as anchor.	5
	Number of anchors	Multiple anchors should be used to measure MID, rather than single anchor.	10
	Measurement properties of anchor	Satisfactory anchor measurement properties (validity, reliability and responsiveness) are the prerequisite of an optimal anchor.	35
PROM issue	Measurement properties of PROM	Required measurement properties for PROM when measuring MID: reliability, validity and responsiveness.	4
		Ordinal scales do not support the mathematical calculations required for MCID calculation.	5
Generalizability of MID	MID varying by contexts	MID may vary depending on patient characteristics.	86
		MID may vary depending on follow-up length.	15
		MID may vary depending on intervention.	28
	MID direction	Distinguishing MID for improvement vs deterioration.	31
	Generic instruments	For generic instruments, MID should not vary depending on the contexts.	3
Statistical issues	Precision of MID	Precision of MID	19
	MID in relation to measurement error	MID should be beyond measurement error.	16
	Triangulation	Triangulation of distribution-based and anchor-based methods.	15
	Representative sample	Representative sample	14
	Relative or absolute MID	Expressing MID in relative or absolute terms †	19
	Analytical method	Possible anchor-based MID analytical methods †	78
		MID may vary depending on the analytical method used.	33
Others	MID selection	MID selection: match study context	8
		MID selection: evidence-based and consensus processes	6
		MID selection: MID levels	1
	A single MID or range of plausible MIDs	A single value or range of plausible MIDs is appropriate for many PROMs	29

Note: * see Appendix 4 for best example of quotation for the items.
† Apply to anchor-based MID reporting only.
ξ Number of citations
MID=minimal important difference; PROM= patient-reported outcome measures.

We found the following issues of particular note. The quotations under item 'MID definition' indicated that most articles that offered a definition (65 out of 95) of MID as 'the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful' [10], representing a modification of the original definition by deleting the phrase 'in the absence of troublesome side-effects and excessive cost, a change in patient management' [1].

Possible anchors included objective clinical measurement, other validated PROMs, global rating of change (also referred as transition scales), and combinations of different anchors. Three items related to the perspective of anchor: that is, who should respond to the anchor question. Two items suggested relying on the perspectives of patients or informed proxies to measure MID, while one item recommended using clinical evaluation (*Table 2*).

The items also reflected methodological criteria for a credible MID: 1) the choice of what constitutes an MID on the anchor should reflect a small but important difference; 2) anchor should be easily understandable; 3) at least moderate correlation between PROM and anchor, which was reported to be at least 0.3 [11] and as high as 0.7 [12]; and 4) precise MID estimates (*Table 2*).

When using transition scales as anchors, additional requirements identified by authors included: 1) relatively shorter recall time, up to 4 weeks [13-15]; 2) transition scale should correlate equally strongly with the pretest and posttest PROM scores but with opposite signs [13]; and 3) the correlation of the transition scale with change scores on PROM should be at least 0.2 greater (in absolute terms) than the correlations with either the baseline or follow-up test scores [13,16] (*Table*

2). Some items captured other credibility criteria, including the same or closely related constructs between anchor and PROM, use of multiple anchors, and satisfactory measurement properties of the anchor (*Table 2*).

When applying an MID to a particular healthcare condition, authors often pointed out that that MIDs may vary across contexts. Of the eligible articles, 91 identified contextual factors including patient characteristics (i.e., age, sex, baseline score, socioeconomic status, locations, disease or conditions), follow-up length, and intervention (*Table 2, Appendix 4*). Authors frequently commented on methodological issues that may influence the application of MID. These included, the direction of MIDs (improvement or deterioration), and the analytical method used to establish MIDs (*Table 2*).

Authors of 15 articles explicitly addressed the selection of optimal MIDs from a candidate pool (*Table 2*). Eight articles recommended that researchers should choose an MID established from a same or similar intervention and patient population to the one to which the MID would be applied. Six articles suggested that the selection should involve expert consensus after review of all available estimates. Based on the methodological rigor of MID generation process, one article categorized MIDs into four levels and suggested using the most rigorous.

Other suggestions for choosing optimal MIDs from a larger pool proved less specific. For example, some authors argued that the MID should be triangulated from distribution-based and anchor-based MID estimates, but without specifying systematic methods for such triangulation (*Table 2*). Similarly, fourteen publications advocated recruiting a representative patient sample to estimate

the MID without providing suggestions on what constitutes a representative sample (*Table 2, Appendix 4*).

4. Discussion

The current literature presents crucial issues potentially relevant for reporting studies estimating anchor-based MIDs and selecting optimal MIDs. The systematic survey identified 34 items related to reporting MIDs, of which 29 items were also related to selecting MIDs. The items summarized quotations from the articles that addressed issues of MID definition and terminology, anchor, PROM, generalizability of MID and statistical issues (*Table 2*). The research team identified at least one direct quotation to support the generation of each item, with a median of 18 quotations per item.

Selecting an MID for use should prioritize MIDs with the greatest credibility [17]. This highlights the need for methodological credibility criteria for anchor-based MID, which investigators could use not only to filter out invalid MIDs but also to guide optimal conduct and reporting of MID studies. Authors of eligible articles suggested credibility issues similar to items in a recently developed anchor-based MID credibility assessment tool [8]*(Table 2*).

Beyond the items in that instrument, the systematic survey identified three potentially additional factors that may influence MID credibility: number of anchors used, measurement properties of an anchor, and construct proximity of an anchor to the target PROM. Using multiple anchors may not, however, be necessary, especially when they are all supposed to lead to one common MID. That is, one anchor, if suitable, may be sufficient.

Further, the suitability of an anchor may be reflected in its correlation with the scores on the PROM under consideration. Indeed, satisfactory correlations may be both the best way of assessing construct proximity and the key measurement property of the anchor for the purpose of MID estimation. However, in the absence of reported correlations – most available MID studies omit this information [7]– subjective assessments of construct proximity may be necessary [18].

The systematic survey identified three potential factors for selecting MIDs: contextual factors, MID direction and analytical method used to estimate MID. However, the importance of each of these factors in explaining MID variability remains uncertain. Exploration of this issue in an established MID inventory [7] may prove enlightening.

Strengths of the systematic survey include the comprehensive search, transparent eligibility criteria, and a protocol detailing criteria and methods for data abstraction. To minimize the possibility of missing relevant articles, and subsequently missing relevant quotations from eligible articles, all reviewers, before screening and data abstraction, underwent calibration exercises. More experienced reviewers resolved any subsequent disagreements. We developed codes to help with quotation abstraction and dynamically added new codes, as necessary, during abstraction. Two investigators (GG, TD) adjudicated challenging discrepancies and uncertainties during data abstraction and, to further ensure the accuracy of final codes, a senior investigator (GG) supervised the cleaning of abstracted codes and quotations. Our steering committee provided oversight for the whole process of the systematic survey.

The systematic survey has limitations. The process of coding, item development as well as code grouping introduced subjectivity, and we did not formally measure agreement between reviewers. Other reviewers may have generated a different list of code and items. Similarly, eligibility decisions involved subjective judgment. The detailed development of criteria, the duplicate decision-making process at each step, and the extensive oversight and checking, may mitigate these concerns. It is also possible that we missed some publications. However, the code system proved saturated after abstraction of 30 eligible publications, reducing the likelihood that we missed key codes and items. Some readers may be interested in the extent to which authors addressed issues in empirical studies versus reviews and commentaries. We did not classify codes in this way, and so cannot provide this information.

The systematic survey used qualitative synthesis to identify methodological issues regarding reporting and selecting anchor-based MIDs. These items may serve as a starting point for further development of a reporting standard for anchor-based MID studies, as well as a systematic selection approach for optimal MIDs. The purpose of the systematic survey was to summarize what has been reported or discussed in current literature: distinguishing suitable from unsuitable items is beyond the scope of this study. Thus, the identified items do not necessarily reflect the best practice for reporting or selecting MIDs – indeed, many items represent mutually exclusive approaches. Therefore, when future investigators refer to the items, they will not adopt all the items for developing a reporting guideline or selecting approaches. Further discussion around what are appropriate items may well be warranted and an ultimate consensus in the expert community would undoubtedly be helpful.

ARTICLE INFORMATION

Author Contributions: GHG, TD, YW conceived the study idea. ACL, BT, CT, GHG, MK, MW, TD, TF and YW developed the protocol and had oversight of the review process. AN, LECL, MB, MG, MP, MRP, QH, TD, YX, YW, YW, VW conducted screening for eligible articles. TD, YW created the data abstraction form and the initial code list for data abstraction. AQ, MB, MP, QH, YW, VW extracted data and conducted coding. GHG, YW developed items. YW wrote the first draft of the manuscript. ACL, BT, CT, GHG, MK, MW, TD, TF and YW interpreted the data analysis, critically commented on the manuscript drafts. ACL, AN, AQ, BT, CT, GHG, LECL, MB, MK, MG, MP, MRP, MW, QH, TD, TF, YX, YW, YW, VW reviewed, revised and approved this manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflict of Interest Disclosures: The authors declared no conflicts of interest.

Funding/Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical approval statement: Not required.

ACKNOWLEDGEMENTS

For this work, we have held several meetings for the steering committee members. It was difficult to find a best time for every member because the time differences. Most of the time, some members had to meet in the early mornings and others needed to stay up late. We really appreciate the flexibility of the steering committee. We also appreciate our reviewers for their voluntary contribution to the work.

Reference

1. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10(4):407-15. doi: 10.1016/0197-2456(89)90005-6
2. Schünemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res* 2005;40(2):593-7. doi: 10.1111/j.1475-6773.2005.00374.x
3. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003;56(5):395-407. doi: 10.1016/s0895-4356(03)00044-1
4. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *Jama* 2014;312(13):1342-3. doi: 10.1001/jama.2014.13128
5. Turner D, Schünemann HJ, Griffith LE, et al. The minimal detectable change cannot reliably replace the minimal important difference. *J Clin Epidemiol* 2010;63(1):28-36. doi: 10.1016/j.jclinepi.2009.01.024
6. de Vet HC, Terwee CB, Ostelo RW, et al. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* 2006;4:54. doi: 10.1186/1477-7525-4-54
7. Carrasco-Labra A, Devji T, Qasim A, et al. Minimal important difference estimates for patient-reported outcomes: A systematic survey. *J Clin Epidemiol* 2020;133:61-71. doi: 10.1016/j.jclinepi.2020.11.024
8. 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. doi: 10.1136/bmj.m1714
9. Devji T, Carrasco-Labra A, Guyatt G. Mind the methods of determining minimal important differences: three critical issues to consider. *Evid Based Ment Health* 2020 doi: 10.1136/ebmental-2020-300164
10. Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77(4):371-83. doi: 10.4065/77.4.371
11. Wu X, Liu J, Tanadini LG, et al. Challenges for defining minimal clinically important difference (MCID) after spinal cord injury. *Spinal Cord* 2015;53(2):84-91. doi: 10.1038/sc.2014.232
12. Malec JF, Ketchum JM. A Standard Method for Determining the Minimal Clinically Important Difference for Rehabilitation Measures. *Arch Phys Med Rehabil* 2020;101(6):1090-94. doi: 10.1016/j.apmr.2019.12.008
13. Guyatt GH, Norman GR, Juniper EF, et al. A critical look at transition ratings. *J Clin Epidemiol* 2002;55(9):900-8. doi: 10.1016/s0895-4356(02)00435-3
14. Kamper SJ, Ostelo RW, Knol DL, et al. 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(7):760-66.e1. doi: 10.1016/j.jclinepi.2009.09.009
15. Lurie F, Kistner RL. In prospective study using Specific Quality of Life & Outcomes Response-Venous (SQOR-V) questionnaire the recall bias had the same magnitude as the minimally important difference. *Qual Life Res* 2011;20(10):1589-93. doi: 10.1007/s11136-011-9910-y
16. 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(4):374-9. doi: 10.1016/j.jclinepi.2008.07.009
17. Koynova D, Lühmann R, Fischer R. A Framework for Managing the Minimal Clinically Important Difference in Clinical Trials. *Ther Innov Regul Sci* 2013;47(4):447-54. doi: 10.1177/2168479013487541
18. Revicki DA, Cella D, Hays RD, et al. Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes* 2006;4:70. doi: 10.1186/1477-7525-4-70

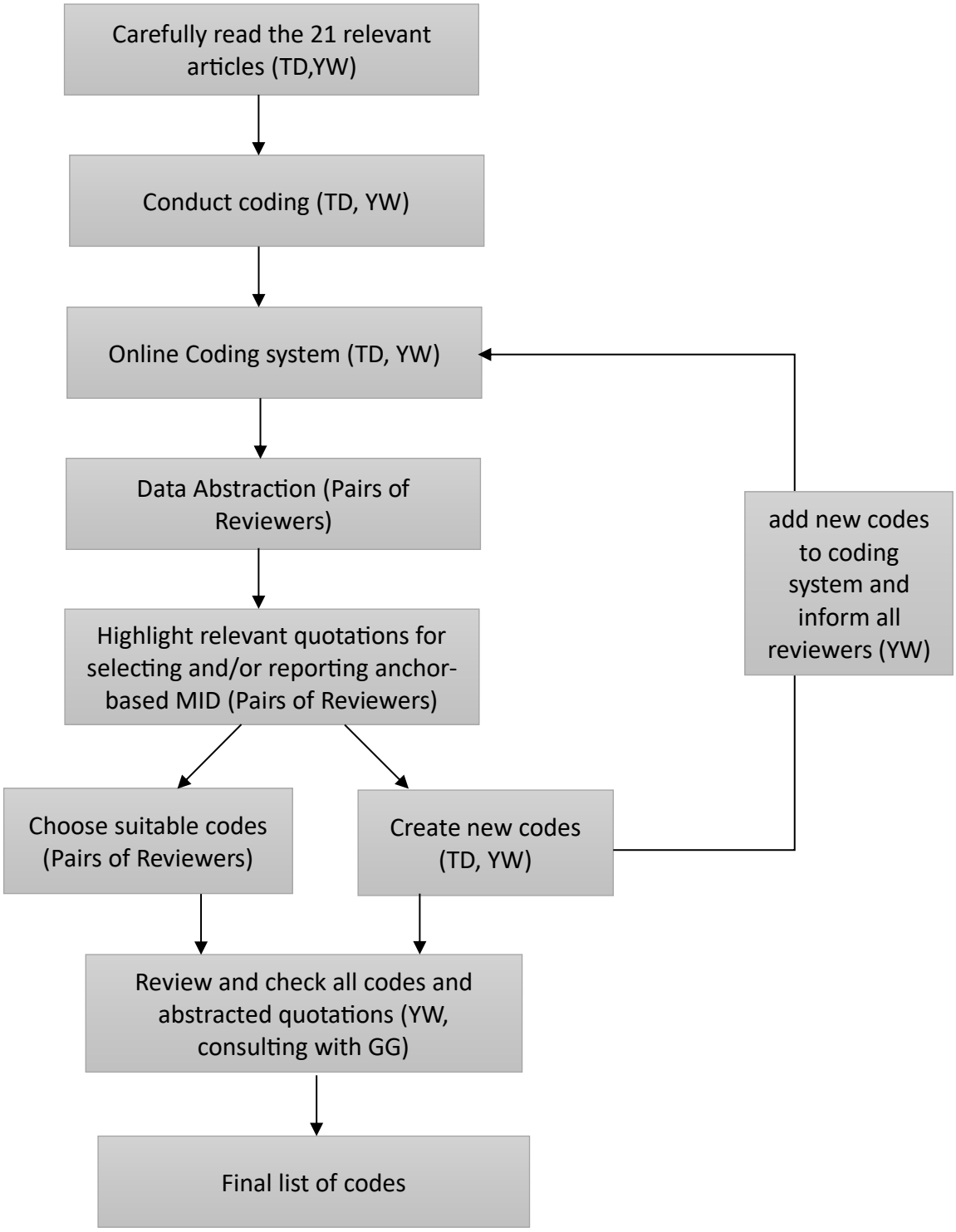
Appendix 1. Searching strategies

Embase/Medline/PsycINFO

1. (clinical* important difference? or clinical* important improvement? or clinical* important deterioration? or clinical* important worsening? or clinical* meaningful difference? or clinical* meaningful change? or clinical* meaningful improvement? or clinical* meaningful deterioration? or clinical* meaningful worsening? or clinical* relevant mean difference? or clinical* significant change? or clinical* significant difference? or mcid or minim* clinical* important or minim* clinical* detectable or minim* clinical* significant or minim* important change? or minim* important difference?).tw.

2. limit 1 to yr="1987 -March.2020 "

Appendix 2. Workflow of Data abstraction and code development.



Appendix 3. The refinement of the initial 63 codes.

Cateogry	Initial code (n=63)	Final Code (n=24)
Definition	MID reporting[1]	MID terminology
	MID definition	MID definition
Anchor issues `	Anchor criteria[2]	/
	GROC[3]	/
	Anchor type	Anchor type
	Anchor criteria: perspective	Anchor perspective
	GROC: patient perspective	Anchor perspective
	Anchor criteria: relevance to patients	Anchor perspective
	Anchor criteria: clinical relevance	Anchor perspective
	Anchor criteria: address importance of change	Anchor threshold
	Anchor criteria: choice of what constitutes an MID on the anchor should reflect a small but important difference	Anchor threshold
	Anchor criteria: interpretability	Anchor interpretability
	GROC: easily understandable	Anchor interpretability
	Correlation between anchor and PROM	Correlation between PROM and anchor
	Correlation between anchor and PROM: GROC	Correlation between PROM and anchor
	Anchor criteria: construct	Anchor construct
	GROC: construct	Anchor construct
	GROC: recall bias	Recall time of transition scale
	Response shift	Recall time of transition scale
	Anchor criteria: single vs multiple anchors	Number of anchors
	GROC: PROM score in unchanged group	Unchanged group of transition scale

	Anchor criteria: measurement properties (i.e. validity, reliability, responsiveness)	Measurement properties of anchor
	GROC: reliability	Measurement properties of anchor
	GROC: validity	Measurement properties of anchor
	GROC: responsive to change	Measurement properties of anchor
PROM issue	Established PROM measurement properties	Measurement properties of PROM
Generalizability of MID	MID variability: age	MID varying by contexts
	MID variability: baseline health status	MID varying by contexts
	MID variability: disease or condition	MID varying by contexts
	MID variability: nationality	MID varying by contexts
	MID variability: patient characteristics	MID varying by contexts
	MID variability: sex	MID varying by contexts
	MID variability: socioeconomic status	MID varying by contexts
	MID variability: follow-up duration	MID varying by contexts
	MID variability: intervention	MID varying by contexts
	MID for improvement vs deterioration	MID direction
	MID variability[4]	Generic instruments
	MID variability: anchor[5]	/
	MID variability: PROM domain[6]	/
Statistical issues	Precision of MID	Precision of MID
	Precision of MID: sample size	Precision of MID
	MID in relation to measurement error of PROM scores	MID in relation to measurement error
	Triangulation	Triangulation
	Representative patient sample	Representative sample
	Expressing MID in relative or absolute terms	Relative or absolute MID

	MID variability: analytical method	Analytical method
	MID calculation method	Analytical method
	Mean change: method	Analytical method
	Mean change: criteria	Analytical method
	Mean change: limitation/strength	Analytical method
	Mean difference: method	Analytical method
	Mean difference: criteria	Analytical method
	Mean difference: limitation/strength	Analytical method
	Regression: method	Analytical method
	Regression: criteria	Analytical method
	Regression: limitation/strength	Analytical method
	ROC: method	Analytical method
	ROC: criteria	Analytical method
	ROC: limitation/strength	Analytical method
Others	MID selection: evidence-based and consensus processes[7]	MID selection
	Single MID or range of plausible MIDs	A single MID or range of plausible MIDs
	Between or within patient approaches conceptual issues[8]	/
	Group or individual level MID[9]	/

Notes: 1. The initial code 'MID reporting' was used as a broad code to capture any general issues regarding MID reporting elements. When refining the code, we deleted irrelevant quotations, assigned relevant quotations to existing codes and created codes for new issues. Finally, we had 6 quotations to support the creation of a new code 'MID terminology'.

2. The initial code 'anchor criteria' was used as a broad code to capture quotations related anchor when reviewers were not very sure which specific anchor code to assign. When refining the code, we deleted irrelevant quotations and assigned relevant quotations to existing codes. Finally, this code was dropped.

3. The initial code 'GROC' was used as a broad code to capture quotations related to GROC anchor when reviewers were not very sure which specific anchor code to assign. When refining the code, we deleted irrelevant quotations and assigned relevant quotations to existing codes. Finally, this code was dropped.

4. The initial code 'MID variability' was used as a broad code to capture quotations related MID variability when reviewers were not very sure which specific MID variability code to assign. When refining the codes, we deleted irrelevant quotations, assigned relevant quotations to existing codes and created codes for new issues. Finally, the code MID variability was dropped, and we had 3 quotations to support the creation of a new code 'generic instruments'.

5. We dropped this code because the issues about anchor have been captured by other anchor codes.

6. We dropped this code because this is irrelevant since it is reasonable to consider different PROM domains as different measurements.

7. The initial code included 3 items: MID selection: match study context; MID selection: evidence-based and consensus processes; and MID selection: MID levels. We finally decided to use a broad code ‘MID selection’ for all instead.
8. We finally considered that even though the initial code captured important MID methodological issues but is irrelevant for either selecting or reporting MID.
9. We finally considered that even though the initial code captured important MID methodological issues but is irrelevant for either selecting or reporting MID.

Appendix 4. Best example of quotation for the items.

Category	Code	Item	Best Example of quotation	N ξ
Definition	MID definition	MID definition †	We now define the MID as the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and which would lead the patient or clinician to consider a change in the management [1].	95
	MID terminology	Standardize MID terminology †	Several terminologies were used to identify the MID. However, using a standardized term and acronym referring to the MID should be investigated in future studies to avoid variability in terminology and to obtain more accurately the maximum number of articles needed for future analysis [2].	6
Anchor issues	Anchor type	Possible anchors†	These criteria, or anchors, may be clinical endpoints (i.e., laboratory measures, physiological measures, clinician ratings), patient-rated global improvement or other PROs with established responsiveness, or some combination of clinical and patient-based outcome [3].	30
	Anchor perspective	The change being perceived as minimally important on the anchor would be perceived as beneficial from the patient's viewpoint.	Research on the interpretability of HRQL instruments should focus primarily on the patient's view...If one accepts that HRQL measurement must be fundamentally patient-centered, the first choice for establishing the MID should be a patient-based approach...Thus, readers, when they interpret the results of research on the MID, must attend to who rated the importance of an HRQL change ... [1]	30
		Proxy perspective is acceptable if patients cannot make decisions.	To further qualify this definition of the MID, only if the MID for informed patients is unknown, if informed patients cannot make decisions about the management of their disease, or if patients prefer informed proxies to make these decisions would one consider the MID estimates of informed proxies [4].	4
		Estimating MID should use clinical evaluation as anchor.	Clinicians also may be appropriate candidates to provide an external assessment of patient change, although without proper training and rigor in making judgments about change, large variations may occur [5].	19
	Anchor threshold	The choice of what constitutes an MID on the anchor should reflect a small but important difference .	Once an anchor is determined, someone then decides what parts of that anchor will be considered "important change" or "small and important change" [6].	38
	Anchor interpretability	Anchor should be easily understandable.	An appropriate anchor should be both interpretable and appreciably correlated with QOL change [7].	18
	Correlation between PROM and anchor	At least moderate correlation between PROM and anchor.	Anchor-based methods require at least moderate correlation of the change on the anchor with the change on the target instrument [8].	36
		Transition scale should correlate equally strongly with the pretest and posttest PROM scores but with opposite signs.	If the variability of the pre and post test scores are equal, were transition measures working in the way they should, one would anticipate an equal and opposite correlation of the transition measure with the pretest score and the post test score [9].	6

		The correlation of transition scale with change scores on PROM should be at least 0.2 greater (in absolute terms) than the correlations with either the baseline or the follow-up test scores.	There should be a negative correlation between the GRC and the baseline instrument score...there should be a positive correlation between the GRC and the follow-up instrument score...and the correlation of the GRC with the difference between follow-up and baseline score should be at least 0.2 greater (in absolute terms) than the correlations with either the baseline or the follow-up test scores [10].	4
	Anchor construct	Anchor and PROMs should measure the same or closely related constructs.	If an anchor-based external criterion is to be used, then the criterion needs to be an independent measure of the same construct, not just another related self-report measure [11].	20
	Recall time of transition scale	The recall time should not be too long when using transition scale as anchor.	It seems clear, therefore, that patients can sometimes recall their prior state, that this recollection bears on their rating of change, and there will be instances when they can retain this memory for periods of up to 4 weeks in duration [9].	53
	Unchanged group of transition scale	The use of unchanged group when using transition scale as anchor.	If it turns out that the change for the no-change group is similar to that of the minimally changed group, then the MID estimate is suspect. However, if the MID change exceeds that of the no-change group, the MID estimate is useful and does not need to be adjusted by the HRQOL change observed in the no-change group [12].	5
	Number of anchors	Multiple anchors should be used to measure MID, rather than single anchor.	This also means that several anchors are needed to accurately assess the MID and to check the robustness and complementarity of the results obtained using different anchors [12].	10
	Measurement properties of anchor	Satisfactory anchor measurement properties (validity, reliability and responsiveness) are the prerequisite of an optimal anchor.	The anchor's validity and reliability are crucial for determination of a valid MCID [13].	35
PROM issue	Measurement properties of PROM	Required measurement properties for PROM when measuring MID: reliability, validity and responsiveness.	For an outcome measure to be clinically useful, the measure must first reflect sound psychometric proper ties of reliability and validity. Beyond this, the outcome tool must demonstrate an ability to accurately detect change, otherwise known as responsiveness [14].	4
		Ordinal scales do not support the mathematical calculations required for MCID calculation.	Measures with an interval level of scaling are required because with such measures the change score will indicate the same degree of change regardless of the initial level on the measure. Ordinal measures can be transformed to interval scaling through Rasch or other item response theory procedures [15].	5
Generalizability of MID	MID varying by contexts	MID may vary depending on patient characteristics.	Clinicians must be aware that MID values for a measure have limited generalizability, meaning that they may not transfer to different patient population (injury type, severity, sex, athletic level, etc) [16].	86
		MID may vary depending on follow-up length.	Differences in MCID in previously reported studies can also be explained by duration of follow-up [17].	15
		MID may vary depending on intervention.	Researchers who incorporate minimally important differences into their methods or evidence users who incorporate minimally important differences into their judgments	28

			should be careful to select a minimally important difference that was derived from a similar cohort of patients who underwent a similar intervention [18].	
	MID direction	Distinguishing MID for improvement vs deterioration.	It is possible that the same change in QOL score on a target instrument warrants different interpretation if it is an improvement, rather than a deterioration [19].	31
	Generic instruments	For generic instruments, MID should not vary depending on the contexts.	..., as a generic instrument measures the generic health status, its MID should not vary across different populations and contexts…[20]	3
Statistical issues	Precision of MID	Precision of MID	Anchor-based approaches in particular suffer from imprecision due to small sample sizes, as this approach uses only a part of all data to estimate the MCID [21].	19
	MID in relation to measurement error	MID should be beyond measurement error.	It is our opinion that any sound MCID value should fulfill two criteria: it has to be at least greater than the measurement error and it has to correspond to the patient perception of importance of change [22].	16
	Triangulation	Triangulation of distribution-based and anchor-based methods.	Early descriptions recommended using multiple methods to determine the MCID and then "triangulating" on the best value. However, a specific or systematic method for this triangulation has not been suggested [15].	15
	Representative sample	Representative sample	To be maximally informative, representative samples of informed patients or if necessary, their proxies should provide estimates of the MID [8].	14
	Relative or absolute MID	Expressing MID in relative or absolute terms †	We found no evidence to suggest consistent or convincing superiority of the relative or absolute approaches [23].	19
	Analytical method	Possible anchor-based MID analytical methods †	The most direct method is to simply calculate the mean PROM score change for patients reporting the transition rating corresponding to the anchor question MCID and equate that value to the PROM's MCID [24].	78
		MID may vary depending on the analytical method used.	There is no consensus in the literature on the most appropriate technique for determining the MCIC; different methods to estimate the MCIC result in different values [25].	33
Others	MID selection	MID selection: match study context	Researchers using previously established MCID cutoff values when setting up new trials are also strongly discouraged if the intervention, outcome measure and patient population are not similar to the settings in which the MCID cut-off was established [26].	8
		MID selection: evidence-based and consensus processes	However, there is a variability among the different estimates from different studies and the choice of the value to be used subsequently is difficult. It should be data-driven and expert-based to have a good face validity [27].	6
		MID selection: MID levels	The 4-level approach for classifying the MCID evidence is summarized…level 1. MCID following specific methodological requirements; level 2. MCID deduced from evidence-based data or RCTs; level 3. MCID deduced from an evaluation of available general databased or methods; level 4. MCID defined by consensus panels or individual opinions...Ideally, level 1 MCID evidence should be used during all stages of a drug development program [28].	1
	A single MID or range of plausible MIDs	A single value or range of plausible MIDs is appropriate for many PROMs	There was debate about whether MICs should be expressed as a single value or as a range that includes all reasonable values. Ranges, however, require the user to know when to use the larger or smaller values. Many may be tempted to use the smallest MIC in order to demonstrate more improvement, but that may not be most appropriate to the patient group or intervention [29].	29

Note: † apply to anchor-based MID reporting only. ξ Number of citations. MID=minimal important difference; PROM= patient-reported outcome measures.

Reference

1. Schünemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res* 2005;40(2):593-7. doi: 10.1111/j.1475-6773.2005.00374.x
2. 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 Qual Life Outcomes* 2018;16(1):228. doi: 10.1186/s12955-018-1055-z
3. Revicki DA, Cella D, Hays RD, et al. Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes* 2006;4:70. doi: 10.1186/1477-7525-4-70
4. Schünemann HJ, Puhan M, Goldstein R, et al. Measurement properties and interpretability of the Chronic respiratory disease questionnaire (CRQ). *Copd* 2005;2(1):81-9. doi: 10.1081/copd-200050651
5. Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. *Phys Ther* 2006;86(5):735-43.
6. Engel L, Beaton DE, Touma Z. Minimal Clinically Important Difference: A Review of Outcome Measure Score Interpretation. *Rheum Dis Clin North Am* 2018;44(2):177-88. doi: 10.1016/j.rdc.2018.01.011
7. Wyrwich KW, Bullinger M, Aaronson N, et al. Estimating clinically significant differences in quality of life outcomes. *Qual Life Res* 2005;14(2):285-95. doi: 10.1007/s11136-004-0705-2
8. Brozek 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
9. Guyatt GH, Norman GR, Juniper EF, et al. A critical look at transition ratings. *J Clin Epidemiol* 2002;55(9):900-8. doi: 10.1016/s0895-4356(02)00435-3
10. 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(4):374-9. doi: 10.1016/j.jclinepi.2008.07.009
11. 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(5):387-97. doi: 10.1097/AJP.0b013e3182327f20
12. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *Copd* 2005;2(1):63-7. doi: 10.1081/copd-200050663
13. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *Jama* 2014;312(13):1342-3. doi: 10.1001/jama.2014.13128
14. Wright A, Hannon J, Hegedus EJ, et al. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). *J Man Manip Ther* 2012;20(3):160-6. doi: 10.1179/2042618612y.0000000001
15. Malec JF, Ketchum JM. A Standard Method for Determining the Minimal Clinically Important Difference for Rehabilitation Measures. *Arch Phys Med Rehabil* 2020;101(6):1090-94. doi: 10.1016/j.apmr.2019.12.008
16. Riemann BL, Lininger MR. Statistical Primer for Athletic Trainers: The Essentials of Understanding Measures of Reliability and Minimal Important Change. *J Athl Train* 2018;53(1):98-103. doi: 10.4085/1062-6050-503-16
17. De Kleermaeker F, Boogaarts HD, Meulstee J, et al. Minimal clinically important difference for the Boston Carpal Tunnel Questionnaire: new insights and review of literature. *J Hand Surg Eur Vol* 2019;44(3):283-89. doi: 10.1177/1753193418812616
18. Jevsevar DS, Sanders J, Bozic KJ, et al. An Introduction to Clinical Significance in Orthopaedic Outcomes Research. *JBJS Rev* 2015;3(5) doi: 10.2106/jbjs.Rvw.N.00064
19. Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77(4):371-83. doi: 10.4065/77.4.371
20. Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life-a systematic review. *J Clin Epidemiol* 2017;89:188-98. doi: 10.1016/j.jclinepi.2017.06.009
21. Keurentjes JC, Van Tol FR, Fiocco M, et al. Minimal clinically important differences in health-related quality of life after total hip or knee replacement: A systematic review. *Bone Joint Res* 2012;1(5):71-7. doi: 10.1302/2046-3758.15.2000065
22. Copay AG, Glassman SD, Subach BR, et al. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. *Spine J* 2008;8(6):968-74. doi: 10.1016/j.spinee.2007.11.006
23. Zhang Y, Zhang S, Thabane L, et al. Although not consistently superior, the absolute approach to framing the minimally important difference has advantages over the relative approach. *J Clin Epidemiol* 2015;68(8):888-94. doi: 10.1016/j.jclinepi.2015.02.017

24. Sedaghat AR. Understanding the Minimal Clinically Important Difference (MCID) of Patient-Reported Outcome Measures. *Otolaryngol Head Neck Surg* 2019;161(4):551-60. doi: 10.1177/0194599819852604
25. Ostelo RW, de Vet HC. Clinically important outcomes in low back pain. *Best Pract Res Clin Rheumatol* 2005;19(4):593-607. doi: 10.1016/j.berh.2005.03.003
26. Draak THP, de Greef BTA, Faber CG, et al. The minimum clinically important difference: which direction to take. *Eur J Neurol* 2019;26(6):850-55. doi: 10.1111/ene.13941
27. 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(7):725-8. doi: 10.1016/j.jclinepi.2008.09.012
28. Koynova D, Lühmann R, Fischer R. A Framework for Managing the Minimal Clinically Important Difference in Clinical Trials. *Ther Innov Regul Sci* 2013;47(4):447-54. doi: 10.1177/2168479013487541
29. Ostelo RW, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine (Phila Pa 1976)* 2008;33(1):90-4. doi: 10.1097/BRS.0b013e31815e3a10

Chapter 3: An Extension MID Credibility Item Addressing Construct Proximity Is a Reliable Alternative to the Correlation Item

Yuting Wang; Tahira Devji; Alonso Carrasco-Labra; Anila Qasim; Qiukui Hao; Elena Kum; Niveditha Devasenapathy; Madeleine T. King; Berend Terluin; Caroline B Terwee; Michael Walsh; Toshi A. Furukawa; Yasushi Tsujimoto; Gordon H. Guyatt

Published in: J Clin Epidemiol. 2023 Mar 5;157:46-52. doi: 10.1016/j.jclinepi.2023.03.001.

ABSTRACT

Objective: Minimal important difference (MID), the smallest change or difference that patients perceive as important, aids interpretation of change in patient-reported outcome measure (PROM) scores. A credibility instrument that assesses the methodological rigor of an anchor-based MID includes one core item addressing the correlation between the PROM and the anchor. However, the majority of MID studies in the literature fail to report the correlation. To address this issue, we extended the anchor-based MID credibility instrument by adding an item addressing construct proximity as an alternative to the correlation item.

Study design and Setting: Informed by an MID methodological survey, we added an alternative item—a subjective assessment of similarity of the constructs (i.e., construct proximity) between PROM and anchor—to the correlation item and generated principles for the assessment. We sampled 101 MIDs and analyzed the assessments performed by each pair of raters. By calculating weighted Cohen's kappa, we assessed the reliability of the assessments.

Results: Construct proximity assessment is based on the anticipated association between the anchor and PROM constructs: the closer the anticipated association, the higher the rating. Our detailed principles address the most frequently used anchors: transition ratings, measures of satisfaction, other PROMs, and clinical measures. The assessments showed acceptable agreement (weighted kappa 0.74, 95% CI 0.55 to 0.94) between raters.

Conclusion: In the absence of a reported correlation coefficient, construct proximity assessment provides a useful alternative in the credibility assessment of anchor-based MID estimates.

Keywords: anchor-based minimal important difference, credibility assessment, patient- reported outcome measure, correlation, construct proximity, reliability.

Running title: An alternative item of the correlation item for MID credibility assessment

What is new?

Key findings

As an alternative to the correlation item in the existing anchor-based MID credibility instrument, we added a credibility item addressing construct proximity between anchor and PROM. Our judgement principles will facilitate others achieving the reliable construct proximity assessment we demonstrated.

What this adds to what was known?

In the absence of reported correlations between anchor and PROM, the extension of the anchor-based MID credibility instrument by an item addressing construct proximity of anchor and PROM can usefully supplement MID credibility assessment.

What is the implication and what should change now?

When original authors fail to report correlations between the anchor and the PROM, investigators using our credibility instrument can use the alternative item--construct proximity assessment; and assign higher credibility to MIDs generated from studies using the anchor of which the construct is more closely related to the PROM construct. When reporting correlations between the anchor and the PROM becomes uniform practice, the alternative item will no longer be necessary.

1. Introduction

Patient-reported outcomes (PROs), often measured by a variety of patient-reported outcome measures (PROMs), capture the patient perspectives regarding treatment effects and are increasingly adopted in clinical trials [1-3]. Scores on PROMs represent the patients' status regarding constructs under assessment at the time of measurement. Interpreting PROM scores requires understanding whether changes in patients' scores represent trivial, small but important, moderate or large changes [4]. Knowledge of the minimal important difference (MID), the smallest change or difference that patients perceive as important, greatly aids such interpretation [5 6].

Typically, to estimate an MID, investigators correlate the difference in PROM scores to an anchor that is itself easily interpretable [5]. Anchor-based MIDs may, however, carry a variety of methodological flaws; for example, the anchor correlates poorly with the PROM (i.e., very low correlation coefficient), which undermine MID credibility [7]. To help distinguish flawed from trustworthy MIDs, the existing anchor-based MID credibility instrument includes 5 core items to assess anchor-based MID credibility [7]. The assessment involves whether it is the patient who responds to the PROM and anchor; the interpretability and relevance of the anchor to patients; the correlation between the PROM and the anchor; the precision of the MID estimate; and whether the chosen threshold for determining the MID on the anchor reflects a small but important difference [7].

One of the core items, the correlation between the PROM and the anchor, provides crucial information: any anchor, to provide valid information to generate a credible MID, must have at least a moderate correlation with the PROM [8-11]. The higher the correlation, the more confident

the inferences regarding MID credibility, and low correlations, certainly below 0.3 and possibly below 0.5, will not inform credible MID estimates [8 9 11-13]. We previously developed an MID inventory that includes 5324 anchor-based MID estimates from 585 primary studies and applied the credibility instrument to each study [13]. About 66% of the studies, however, failed to report the correlations [14 15].

In a linked systematic survey, we qualitatively synthesized methodological issues related to the reporting of studies estimating anchor-based MIDs [16]. In addition to the items in the original credibility instrument [7], researchers frequently noted that the extent to which the PROM and anchor are measuring the same construct – what we will call construct proximity – could affect the credibility of anchor-based MIDs: MIDs estimated from very closely related constructs of the anchor and PROM (e.g., both measure pain) would likely result in a higher correlation than those estimated from anchors with constructs that differ from the PROM (e.g., physical function limitation vs change in hemoglobin levels) [16]. Therefore, if constructs are similar and other credibility criteria are met, it is likely MIDs will be credible; if constructs are very different, it is unlikely credible MIDs will emerge.

Subsequently, in reviewing the methods of studies estimating MIDs included in the MID inventory [13], we noted a spectrum of construct proximity between the PROM and the anchor: from very similar constructs to egregiously different constructs. When investigators fail to report correlations between the PROM and the anchor, the addition of a subjective assessment of construct proximity between PROM and anchor as an alternative to the correlation item would supplement MID credibility assessment. To implement the assessment, we developed detailed judgement principles,

applied the principles to a sample of MIDs, and determined the inter-rater reliability of the resulting ratings. In this article, we report the methods and results of this effort.

2. Methods

2.1 Development of the wording of the alternative credibility item, response options and judgement principles

Through brainstorming, example-based discussions and iterative refinement of ideas, a core team, consisting of clinicians, health research methodologists and clinical epidemiologists (AC-L, AQ, GHG, TD, and YW) developed the initial draft of the alternative item, the response options, and the judgement principles for construct proximity assessment. We designed judgement principles for different anchor types, chosen on the basis of the frequency with which they appeared in the MID inventory, including transition ratings, measures of satisfaction, other PROMs, and clinical measures [13]. An additional group of investigators (BT, CT, MK, MW, TAF, YT) reviewed suggestions from the core team, and ultimately approved the final response options and judgement principles.

2.2 Inter-rater reliability of the construct proximity credibility item

2.2.1 Data source and rater training

The data were drawn from the MID inventory [14]. The raters' assignment was to judge the extent to which the anchor construct matched that of the PROM. Prior to the assessment, the core team developed detailed instructions with illustrative examples for the judgement principles. Two members of the core team (YW and TD) presented these examples to the raters (AQ, EK, QH, ND) and discussed them thoroughly. After the instruction session, raters completed a calibration

exercise assessing a sample of 13 MID studies. Subsequently, raters, working in pairs, independently applied the judgement principles to evaluate the construct proximity of PROMs and their associated anchors.

2.2.2 Sample size and sampling

Because the main concern for the reliability of the alternative item was the consistency of implementing the judgement principles among raters, we tested inter-rater reliability. We used an expected reliability of 0.7 from a weighted Cohen's kappa with a fixed width of the 95% confidence interval (CI) of 0.2 for sample size calculation. For two raters assessing the same MIDs, achieving this level of precision would require at least 101 MIDs [17].

To ensure independent observations, for the studies with multiple estimates, we selected only the first MID estimate appearing in the inventory. We randomly sampled 101 MID studies from our inventory [14] and analyzed the assessments performed by each pair of raters.

2.2.3 Reliability Analysis

To assess inter-rater reliability, we calculated a weighted Cohen's kappa and the associated 95% CI (http://www.vassarstats.net/kappa.html). Assuming the response options as ordinal and equidistant, calculations used quadratic weights based on the formula: $wi=1-(i^2/(k-1)^2)$, where i is the distance between categories of response options and k is the total number of categories. We combined the response of '*impossible to tell*' and '*definitely not related'* (see results section below) and considered a minimum reliability coefficient of 0.7 as acceptable [6].

3. Results

3.1 Wording of the alternative item and response options

Item 3.1. To what extent is the construct of the anchor closely related to the construct of the PROM?

The following five-point adjectival scale frames the response options, including '*definitely closely related'; 'to a great extent'; 'not so much'; 'definitely not related'; and 'impossible to tell'*. Raters base their judgements on the extent of the similarity and anticipated associations between the PROM construct and the anchor construct: the more similar and closer the anticipated association, the higher the rating. For example, a response of '*definitely closely related'* means the two constructs of PROM and anchor are very similar or very closely related, while the option '*definitely not related*' means that the two constructs are different and unrelated. In the absence of sufficient details to make an informed judgment, raters use a response of '*impossible to tell*'.

3.2 Judgement principles

Below, we provide detailed judgement principles for the most frequently used anchors, including transition ratings, measures of satisfaction, other PROMs, and clinical measures [14], to guide raters to an optimal response option in a systematic and logic way. In *Appendix 5*, we provide worked examples to further explain the application of the judgement principles.

3.2.1 Transition rating anchor

1. If assessors judge the two constructs of the PROM and transition rating anchor as the same or very similar, use a rating of '*definitely closely related*'.

2. When the two constructs refer to different targets of measurement, the rating should be lower. Depending on the likelihood of the two different constructs changing in parallel (i.e., the anticipated association), assessors can choose a rating ranging from '*definitely not related'* to '*to a great extent'*.

3. Often, the anchor construct description is vague, for example 'health status'; the assessors must take into account the clinical condition(s) and any interventions administered, and make inferences regarding how the participants in an MID study would perceive the construct (*Appendix 5, example 1*).

4. Because health states are essentially subjective experiences best assessed by patients rather than by clinicians, the anticipated association between a PROM and a clinician-rated transition anchor would likely be lower than with a patient-rated anchor. Although we have captured 'who responds to PROM and anchor' in another credibility item [6], to infer the likely magnitude of the association between the PROM and the transition anchor for the judgement of construct proximity, we still need to consider who completes the transition ratings. Therefore, after applying the first three principles, assessors should further rate down for clinician-rated transition ratings (*Appendix 5, example 2*).

3.2.2 Anchor measuring patient satisfaction

Some consider patient satisfaction a measure related to quality of health care rather than outcomes. However, when one asks, for instance, "*How satisfied are you with your shoulder function*" [18], there is no reference to the quality of care but only to the outcome. We included such satisfaction measures as PROMs in our inventory. When judging the construct proximity, details regarding the target of the anchor measuring satisfaction (i.e., with exactly 'what' is the respondent satisfied or dissatisfied) is important. As noted below, the anchor may refer to outcomes or quality of care, and if quality of care, there is likely poor construct proximity.

1. When a satisfaction anchor does not specify the satisfaction target (e.g., *are you satisfied?* without further specification), because participants could perceive the satisfaction target

very differently, leading to possibly great variations in the perceived constructs of the anchor, regardless of the PROM construct, raters should choose '*definitely not related'*.

2. For anchors measuring satisfaction and PROMs that also address satisfaction, if both share the same satisfaction target, raters should choose '*definitely closely related*'. If, however, the targets differ (e.g., the anchor asks about satisfaction with quality of care whereas the PROM asks about satisfaction with pain relief), depending on the anticipated association between the two targets, raters can give a rating of '*to a great extent'*, '*not so much*', or '*definitely not related'*.

3. For anchors measuring satisfaction and PROMs that address something other than satisfaction, the construct proximity rating could be, at highest, '*to a great extent'*. Then, details regarding the satisfaction target of the anchor should inform the final rating. Below, as illustrations, we provide two commonly used targets of anchors measuring satisfaction.

3.1 A typical target is the experienced outcome after an intervention. For example, an anchor might ask patients their willingness to have the intervention again given the experienced outcome. By taking into account the clinical condition(s), raters could be able to infer the experienced outcome to which the anchor refers. If it is the same as the construct of the PROM (e.g., pain relief), raters should choose the highest rating of '*to a great extent'*. If different, depending on the anticipated association between the target and PROM construct, raters can assign a rating of '*not so much'* or '*definitely not related'* (*Appendix 5, example 3*).

3.2 Another typical target is care management or satisfaction with an intervention. For example, an anchor might ask patients whether they are satisfied with the care they received. When anchors refer to satisfaction with quality of care, participants

may consider issues, including the experienced outcome, costs, care center environment, satisfaction with the behavior of the care provider. Because these issues may or may not related to the construct the PROM addresses, raters should choose '*not so much*' or '*definitely not related*' (*Appendix 5, example 4*).

3.2.3 Other PROM as an anchor

Knowledge of at least the subdomains and possibly the items of the anchor PROM and the target PROM are necessary to judge the constructs they are addressing and thus make an accurate rating of their construct proximity.

1. For an anchor PROM measuring the same or very similar construct(s) as the target PROM, if the two PROMs are designed for the same condition or disease, raters can choose a highest rating of '*definitely closely related*'. If, however, the two PROMs are designed for different conditions or diseases, because it is likely that even the same constructs might cover different domains and mean different things to patients, raters should choose a highest rating of '*to a great extent*'. Then, raters may use the details of the two PROMs - the summary of subdomains and items - to inform a more accurate rating. When the subdomains and items of the two PROMs address very similar aspects, raters should choose "*to a great extent'*; otherwise, they should rate down further (*Appendix 5, example 5*).

2. For an anchor PROM measuring different construct(s) than the target PROM, a highest possible rating should be '*to a great extent'*. Depending on the likelihood of the two different constructs changing in parallel, or how similar the subdomains (and items) are, assessors may or may not rate down further.

3.2.4 Clinical measure as an anchor

1. The ratings would depend on the biological relation between the clinical measures (e.g., physiologic measures) and the PROM construct (*Appendix 5, example 6*).

2. The clinical conditions and interventions administered in an MID study may affect the biological relation. Assessors may also need clinical input to infer the biological relation.

3. As clinical measure tends to be poorly correlated with PROs [1 2], assessors should seldom choose '*definitely closely related*'.

3.2.5 Multiple anchors

If investigators have used multiple anchors to estimate an MID, unless investigators specify, there is no assurance that they have put greater weight on one anchor or other; and raters should assume that they weighted all anchors equally. Raters should then give a final rating lying between the ratings for the most related constructs of the PROM and anchor and the least related. When a majority rating exists (e.g., two out of three anchors with a rating of '*not so much'*), raters should choose the rating of the majority as the final rating.

3.3 Reliability analysis

Three pairs of reviewers (ND and TD, 30 assessments; EK and YW, 34 assessments; and AQ and QH, 37 assessments) conducted the construct proximity assessment. The sample anchors included transition ratings (N=66), satisfaction measures (N=8), PROMs (N=7), clinical measures (N=6), multiple anchors (N=3), and others (clinician-reported outcome measure (N=3), age group (N=1), disease related outcome (N=4), and comparison to another patients (N=3)). The inter-rater reliability was 0.74, with a 95% CI of 0.55 to 0.94 (*Table 3*).

Table 3. Concordance in the construct proximity assessment between raters.

N=101 MIDs		Rater 1				
		Definitely not related/impossible to tell	*Not so much*	*To a great extent*	*Definitely closely related*	Weighted kappa (95%CI)
Rater 2	*Definitely not related/impossible to tell*	4	0	0	0	-
	Not so much	0	11	4	0	-
	To a great extent	0	8	29	8	-
	Definitely closely related	0	1	11	25	-
						0.74 (0.55-0.94)

4. Discussion

4.1 Main findings

To address situations in which investigators do not report correlations between the PROM under consideration and an anchor, we developed an alternative item addressing the construct proximity between the PROM and the anchor. We developed detailed principles for making proximity judgments for various types of anchors used by investigators in studies estimating MIDs. Guided by the judgement principles, raters showed acceptable agreement on the assessments (*Table 3*), demonstrating acceptable inter-rater reliability (weighted Cohen's kappa=0.74). Future users can use the judgement principles described here, with reference to examples (*Appendix 5*), to conduct the assessment for this alternative credibility item.

4.2 Strengths and limitations

The paucity of reported correlation coefficients between PROMs and anchors [14], as well as the existence of instances in which the target concept of the anchor bore little relation to the target concept of the PROM, established the need to add an alternative item for assessing credibility. A comprehensive systematic survey on anchor-based MID methodology [16] informed the suitability of assessing construct proximity as a substitute for the correlation coefficient between the PROM and the anchor. The team developing the judgement principles included developers of credibility instruments [6] and the MID inventory [13], and methodologists having rich experience in anchor-based MID estimation. Examples in the MID inventory informed the refinement of judgement principles. The acceptable inter-rater reliability attests to the usefulness of this alternative item.

The major limitation of construct proximity assessment is its subjectivity: inferences regarding anticipated correlations involve rater subjectivity. The assessment may require a level of knowledge about PROM and anchor, which may affect final judgements. The acceptable agreement among our raters, however, indicates that explicit principles and training can result in satisfactory calibration of raters.

4.3 Implications

MIDs aid the interpretation of PROM results in clinical trials. If MID estimates do not, however, actually reflect underlying true MIDs, interpretation of the magnitude of treatment effects will be flawed. Such misleading estimates likely occur when the methodology of MID estimation is seriously flawed. Thus, selecting an MID estimate for clinical trials should consider its credibility. The choice of an anchor plays a crucial role in determining the credibility of MID estimates. The appropriateness of an anchor depends on how well it captures the same construct as the target PROM [12 19-21]. A suitable anchor measures a construct similar to that of the target PROM, but

unsuitable anchor will not. To achieve a trustworthy estimation of anchor-based MIDs, investigators should therefore choose anchors that measure constructs as similar as possible to the target PROM [22].

Construct proximity between the PROM and the anchor is best quantified by correlations. Close anchor-PROM construct proximity does not guarantee a satisfactory correlation between the PROM and the anchor. It therefore remains critical that investigators in future report correlations between PROM and anchor – when this becomes uniform practice, our alternative item will become obsolete. If correlations are, however, not reported, investigators using our credibility instrument should now include the alternative item and assign higher credibility to MIDs generated from more closely related constructs [23 24]. Our judgement principles will facilitate the assessment of the extent to which the two constructs are similar or related.

5. Conclusions

In the absence of reported correlations between the PROM and the anchor, construct proximity assessment represents a useful alternative in the anchor-based MID credibility instrument. The judgement principles facilitate reliable construct proximity assessment for the PROM and the anchor. Although adding construct proximity assessment in the credibility instrument to some extent remedies the failure of investigators to report the correlation coefficients, the only fully satisfactory solution is for investigators to uniformly report the correlations.

Article information

Author Contributions: GHG, TD, TAF, YT, YW conceived the study idea. ACL, AQ, GHG, TD, YW conducted series of discussions for the development of the item and judgement principles. BT, CT, MK, MW, TAF, YT reviewed, added suggestions to and ultimately approved the final version of the item wording, response options and judgement principles. TD and YW conducted the introduction sessions for raters. AQ, EK, QH, ND, TD, YW assessed the samples. YW wrote the first draft of the manuscript. ACL, AQ, BT, CT, EK, GHG, QH, MK, MW, TD, TAF and YT critically commented on the manuscript drafts. ACL, AQ, BT, CT, EK, GHG, QH, MK, MW, TD, TAF, YT and YW reviewed, revised and approved this manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflict of Interest Disclosures: The authors declared no conflicts of interest.

Funding/Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical approval statement: Not required.

Acknowledgement

The authors thank Mark Phillips, Bradley C Johnston, Dena Zeraatkar, Meha Bhatt, Xuejing Jin, Romina Brignardello-Petersen, Olivia Urquhart, Farid Foroutan, Stefan Schandelmaier, Hector Pardo-Hernandez, Robin WM Vernooij, Hsiaomin Huang, Linan Zeng, Yamna Rizwan, Reed Siemieniuk, Lyubov Lytvyn, Zhikang Ye, Liam Yao, Vanessa Wong, Donald L Patrick, Shanil Ebrahim, Gihad Nesrallah, Holger J Schunemann, Mohit Bhandari, and Lehana Thabane for their contributions on the MID inventory and credibility instrument projects. The authors also thanks

Ying Wang, Maryam Ghadimi, Yaping Chang, Layla Bakaa, Sam Al-Rammahy, Mike Ge for their contributions to the abstraction of anchor questions.

Reference

1. Jones PW. Health status measurement in chronic obstructive pulmonary disease. *Thorax* 2001;56(11):880-7. doi: 10.1136/thorax.56.11.880
2. Yohannes AM, Roomi J, Waters K, et al. Quality of life in elderly patients with COPD: measurement and predictive factors. *Respir Med* 1998;92(10):1231-6. doi: 10.1016/s0954-6111(98)90426-7
3. Vodicka E, Kim K, Devine E, et al. Inclusion of patient-reported outcome measures in registered clinical trials: evidence from ClinicalTrials. gov (2007–2013). *Contemporary clinical trials* 2015;43:1-9.
4. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;118(8):622-9. doi: 10.7326/0003-4819-118-8-199304150-00009
5. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10(4):407-15. doi: 10.1016/0197-2456(89)90005-6
6. Schünemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res* 2005;40(2):593-7. doi: 10.1111/j.1475-6773.2005.00374.x
7. 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. doi: 10.1136/bmj.m1714
8. Guyatt GH, Norman GR, Juniper EF, et al. A critical look at transition ratings. *J Clin Epidemiol* 2002;55(9):900-8. doi: 10.1016/s0895-4356(02)00435-3
9. 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(1):37-45. doi: 10.1016/j.jclinepi.2009.03.011
10. Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life-a systematic review. *J Clin Epidemiol* 2017;89:188-98. doi: 10.1016/j.jclinepi.2017.06.009
11. Guyatt GH. Making sense of quality-of-life data. *Med Care* 2000;38(9 Suppl):Ii175-9. doi: 10.1097/00005650-200009002-00027
12. Ward MM, Guthrie LC, Alba M. Domain-specific transition questions demonstrated higher validity than global transition questions as anchors for clinically important improvement. *J Clin Epidemiol* 2015;68(6):655-61. doi: 10.1016/j.jclinepi.2015.01.028
13. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res* 2011;11(2):171-84. doi: 10.1586/erp.11.9
14. 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. doi: 10.1016/j.jclinepi.2020.11.024
15. Devji T, Carrasco-Labra A, Guyatt G. Mind the methods of determining minimal important differences: three critical issues to consider. *Evid Based Ment Health* 2021;24(2):77-81. doi: 10.1136/ebmental-2020-300164
16. Wang Y, Devji T, Qasim A, et al. A systematic survey identified methodological issues in studies estimating anchor-based minimal important differences in patient-reported outcomes. *J Clin Epidemiol* 2021;142:144-51. doi: 10.1016/j.jclinepi.2021.10.028
17. Shoukri MM, Asyali M, Donner A. Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research* 2004;13(4):251-71.
18. Zhou L, Natarajan M, Miller BS, et al. Establishing Minimal Important Differences for the VR-12 and SANE Scores in Patients Following Treatment of Rotator Cuff Tears. *Orthop J Sports Med* 2018;6(7):2325967118782159. doi: 10.1177/2325967118782159
19. Fayers PM, Hays RD. Don't middle your MIDs: regression to the mean shrinks estimates of minimally important differences. *Qual Life Res* 2014;23(1):1-4. doi: 10.1007/s11136-013-0443-4
20. McClimans L. Interpretability, validity, and the minimum important difference. *Theor Med Bioeth* 2011;32(6):389-401. doi: 10.1007/s11017-011-9186-9

21. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *Copd* 2005;2(1):63-7. doi: 10.1081/copd-200050663
22. 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(4):374-9. doi: 10.1016/j.jclinepi.2008.07.009
23. Revicki DA, Cella D, Hays RD, et al. Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes* 2006;4:70. doi: 10.1186/1477-7525-4-70
24. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61(2):102-9. doi: 10.1016/j.jclinepi.2007.03.012

Appendix 5. Worked examples for assessing the construct proximity.

- **Transition ratings**

Example 1:

Investigators calculated the minimal important difference (MID) for the physical function domain of the Western Ontario and McMaster University Osteoarthritis Index (WOMAC) for patients with femoroacetabular impingement undergoing operation and used a transition rating anchor asking patients 'How much did the operation help your hip problem: helped a lot, helped, helped only little, didn't help, made things worse'.

Rating: *to a great extent*

Rational: The two constructs were different: the PROM was 'physical function' and the anchor was 'hip problem'. The rating would thus depend on the likelihood of the two constructs changing in parallel. However, we need to infer how patients would perceive 'hip problem'. Since the patient condition was femoroacetabular impingement, 'hip problem' could be mainly about two aspects: functional limitation and pain. Therefore, the two different constructs, to some extent, were related. We should thus rate it as a '*to a great extent*'.

Example 2:

Investigators calculated the minimal important difference (MID) of VAS-pain for the patients with osteoarthritis and used a transition rating anchor asking **the attending surgeon** to evaluate the overall status of patients following surgery.

Rating: ***not so much***

Rational: Since the **condition** of the included patients was osteoarthritis, we assumed '**overall status**' mainly referring to 'pain and function limitation'. Although the two constructs of the PROM and the anchor were different, we suspected a moderate correlation and rate the construct proximity as a '*to a great extent*'. But the anchor was assessed by the attending surgeon. We should further rate down and give a final rating of **not so much**.

- **Satisfaction anchor**

Example 3:

Investigators calculated the MID for the Pain visual analogue scale (VAS-Pain) for patients with low back pain undergoing surgery and used a satisfaction anchor asking their willingness to have the surgery again given the experienced outcome.

Rating*: to a great extent*

Rational: The PROM was measuring pain but not a satisfaction scale, so that the highest rating was set as '*to a great extent*'. Considering the patient condition was low back pain, we could infer that the patients perceived 'the experienced outcome' after surgery as 'relief of pain'. The anchor was thus about the satisfaction with 'relief of pain'. Because the anchor satisfaction target was same as the PROM construct – 'pain', we rated it as '*to a great extent'*.

Example 4:

Investigators calculated the minimal important difference (MID) for the pain domain of the Western Ontario and McMaster University Osteoarthritis Index (WOMAC) for

patients with Knee osteoarthritis undergoing total knee replacement and used a satisfaction anchor asking 'What is your global level of satisfaction with surgical management?'.
Rating: *not so much*
Rational: The PROM was measuring pain but not a satisfaction scale, so that the highest rating was set as '*to a great extent'*. However, the satisfaction target of the anchor is about 'surgical management'. The patients would refer to quality of care and consider issues including the experienced outcome (may include pain), costs, satisfaction with the behavior of the care provider. In this case, the anchor construct could potentially have no or a low correlation with the PROM construct--pain. We should thus rate down and rate it as '*not so much*'.

- **PROM anchor**

Example 5:
Investigators calculated the MID for St George's Respiratory Questionnaire (SGRQ) for patients with chronic obstructive pulmonary disease and used EuroQol-5 Dimensions (EQ-5D) instrument as the anchor.
Rating: *not so much*
Rational: Both SGRQ and EQ-5D measure quality of life, but the two PROMs were not designed for a same condition. SGRQ is disease-specific instrument designed for patients with obstructive airways disease, while EQ-5D is a generic instrument applicable for all people. Therefore, at most, we should give a rating of '*to a great extent'*. By comparing the summary of subdomains (SGRQ: symptoms, activities, impact; EQ-5D: mobility, self-care, usual activities, pain/discomfort and anxiety/depression), we considered further rating down because few overlaps between the subdomains and rated it as '*not so much'*.

- **Clinical measurement**

Example 6:
Investigators calculated the MID for Perform Questionnaire (PQ)-physical limitation for patients with cancer receiving conservative cancer therapies and used change in hemoglobin (Hb) level as the anchor.
Rating: *definitely not related*
Rational: We considered that the change in Hb level (e.g., a change of 2 g/l) was not relevant to the change in physical function and questioned the existence of biological relation between them. Therefore, a rating of '*definitely not related'* was suitable.

*These examples are from the real examples of the MID inventory [13].

Chapter 4: A Step-by-Step Approach for Selecting an Optimal Minimal Important Difference

Yuting Wang; Tahira Devji; Alonso Carrasco-Labra; Madeleine T. King; Berend Terluin; Caroline B Terwee; Michael Walsh; Toshi A. Furukawa; Gordon H. Guyatt

Published in: BMJ 2023;381:e073822. doi: https://doi.org/10.1136/bmj-2022-073822

Abstract

Researchers proposed the minimal important difference (MID), the smallest change or difference that patients perceive as important, to aid interpretation of patient-reported outcomes measure (PROM) scores. When multiple MIDs for a given PROM differ substantially, selecting an optimal MID to aid interpretation may prove challenging. To address the problem, we developed a systematic, step-by-step selection approach. An optimal MID, at least, should be methodologically sound and secondarily, should, as far as possible, match the intended application contexts. Therefore, this approach is geared to explaining the variability of the MIDs for the PROM of interest by the methodological rigor and contextualized factors influencing the MID application, and where appropriate, provides a single optimal MID, i.e., the median of the selected estimates in a relatively narrow range.

Summary points

- A systematic step-by-step approach has been developed to select an optimal, anchor based, minimal important difference (MID) from various MID estimates of a given patient reported outcomes measure (PROM).
- This approach relies on information on credibility and contextualised factors of all available anchor based MIDs of a target PROM to select the optimal anchor based MID.
- The approach resolves the difficulties in choosing an optimal MID among multiple MIDs to interpret PROM results, which will prove helpful for clinical trials, systematic reviews, and clinical practice guidelines that use PROMs.
- An optimal MID should be credible and should, as far as possible, match the intended application contexts.

Background

When measuring a health state experienced and best known by patients, physiological measures and clinicians' estimates have serious limitations. Thus, directly measuring the patient perspective represents the only satisfactory approach. Clinicians and researchers can measure patient experience - including symptom status, physical function, mental health, social function, wellbeing, and quality of life - using patient reported outcome measures (PROMs). Use of PROMs enhances understanding of the effects of interventions designed to impact disease status and course on patients' lives [1-3]. Authorities thus advocate for using PROMs as endpoint measures in clinical trials examining treatment effects [4-8].

It is challenging, however, to interpret PROM results. Provided large enough sample size, small differences in PROM scores within or between groups that may not be important to patients could achieve statistical significance [9]. Researchers proposed the minimal important difference (MID), the smallest change or difference that patients perceive as important (on average), to aid interpretation of PROM scores [10 11]. The MID has clear implications for interpreting differences between groups in randomized trials: the smaller the mean differences in relation to the MID, the less likely differences represent important and substantial effects, and the larger the difference the more likely. Thus, when presenting results as mean difference in scores between groups, MIDs inform judgements as to whether mean treatment effects represent trivial, small but important, moderate, or large effects [12]. MIDs can also work as a threshold for responder analyses in which trialists estimate the proportion of patients who have achieved an important treatment benefit [13].

Researchers can choose between two approaches for estimating an MID: anchor-based and distribution-based methods. Anchor-based methods examine the relation between a PROM of interest and an anchor that is itself easily interpretable (*Box. Glossary*) [10]. Distribution-based methods use the statistical characteristics of PROM scores to estimate MIDs, thus providing no clear relation to the importance of the change in PROM scores to patients [14]. Anchor-based MIDs therefore represent a far better approach to aid interpretation of the magnitude of treatment effects [14-16].

Conducting a clinical trial or systematic review using PROMs requires a predefined MID at the stage of developing the protocol [5-7]. For a given PROM, however, multiple anchor-based MIDs differing substantially from one another are often available [17-19]. Currently, researchers may assume all MID estimates from published studies that have undergone peer review are equally trustworthy, and randomly choose one for their use [20-22]. Such a practice risks choosing a misleading MID estimate, and thus misinterpretation of results.

In response to this problem, the MID research community has sought approaches to selecting optimal MIDs. Some have suggested using MIDs meeting specific methodological requirements [23]. Others have suggested MID selection be context-dependent and therefore encouraged using an MID established in contexts similar to the trial [15 24 25]. When variability exists among available MID estimates, a consensus process could inform MID selection [26 27], or one might triangulate different MIDs to generate one single MID [28-30]. These suggestions all have limitations including narrowness of perspective, lack of specification and, until recently, lack of criteria to define methodology that will produce a trustworthy MID.

Our team developed an instrument to evaluate the credibility of anchor-based MIDs [31]. The instrument and its recent extension [32] provide a systematic approach to deal with the methodological assessment for anchor-based MIDs. We have also developed a living anchor-based MID inventory (Patient Reported Outcome Minimal Important Difference (PROMID) Database) that includes all available estimates of known PROMs from the literature and includes credibility assessment of each MID estimate (www.promid.org) [33].

The development of the inventory and assessment of credibility did not, however, solve the problem of choosing an optimal MID estimate: when a number of widely varying MID estimates are available, credibility may be similar across estimates. Thus, a systematic approach to selecting the optimal anchor-based MID remains unavailable. To address this deficiency, we have developed an approach. Here, we describe the methods of development, the rationale for the approach, and the steps to select the optimal MID from available MID estimates.

Box. Glossary terms regarding minimal important differences (MIDs)

Anchor-based MID

MID refers to the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful. Anchor-based MIDs relate a difference in the target PROM to an independent measure (i.e., the anchor) that is itself interpretable. For example, investigators use a transition rating scale (i.e., global rating of change) as the anchor (e.g., *Since last month when we started the new treatment, are you feeling better or worse and, if so, to what extent?", with responses of 'very much better, much better, moderately better, slightly better, about the same, slightly worse, moderately worse, much worse, very much worse*') and establish the anchor-based MID for a PROM by estimating the average change in the small but importantly improved group (i.e., MID group) on the anchor (e.g., '*slightly better*' group), where the changes in the PROM score should at least have a moderate correlation (i.e., correlation coefficient=0.5) with the transition rating scale.

The most credible MID estimates

The most credible MID estimates refer to the available MID estimates that receive the highest ratings across the five core credibility criteria (i.e., the highest credibility rank, see *Appendix 8*).

Consistency among MID estimates

Consistency is defined as 80% of the MID estimates lying within an absolute value of 10% of the PROM score above or below the median (i.e., within a range of 20% of the median).

Box. Glossary (continued)

Near the median of the whole distribution

The absolute difference between the median of the selected MID estimates and the median of the whole distribution is less than an absolute value of 10% of the PROM score.

Enough MID estimates

Analogously to subgroup analysis, we will use the same threshold for enough MID estimates. That is, at least 3 estimates per contextualized factor.

Explained MID variability

For binary contextualized factors, we categorize the MID estimates into two groups (e.g., surgical group vs non-surgical group), and to test whether variability is explained by the factors, we compare the medians of the two groups using Wilcoxon Rank Sum Test with a threshold p value of 0.10.

Methods

Following formation of a steering committee, the development of the selection approach consisted of 3 stages: conducting a systematic survey to identify issues related to selecting MID; gathering expert views on the general selection framework; reaching consensus on the details under the selection framework; and formulating a systematic, step-by-step selection approach.

Steering Committee

Prior to the start of the project, we established a steering committee including clinicians, health research methodologists and clinical epidemiologists (AC-L, BT, CBT, GHG, MTK, MW, TD, TAF and YW), a number of whom have rich experience in health status measurement and MID research. The steering committee regularly attended virtual meetings to discuss outstanding issues and decide the next steps of the development. We recorded the meetings and circulated summaries of discussions.

Systematic survey identified candidate items related to MID selection

We conducted a systematic survey, searching up to March 2020, to qualitatively summarize items researchers and methodologists have offered to select optimal MID estimates. We have previously reported a detailed description about the systematic survey [34]. Briefly, the survey identified 29 items that constituted candidate criteria for selecting an optimal MID. They covered MID methodology issues, including anchor, PROM and MID-related statistical issues, as well as the factors impacting the generalizability of MID estimates, including the contexts in which the estimates were developed [34].

Expert views on the selection framework and decisions on the candidate items

In parallel with the systematic survey, we collected the committee's views on the selection framework. Through discussions, the committee agreed on two key broad criteria: MID methodological rigor and generalizability. That is, to develop the selection approach for an optimal MID, the most crucial and first criterion is the methodological rigor of MID estimate development. Secondarily, the optimal MID should, as far as is possible, match the intended application contexts.

Then, we reached consensus on items identified from the survey [34] to be included in the selection framework, which informed the following development of the systematic selection approach. A core team (AC-L, GHG, TD and YW) first conducted intensive discussions on the candidate items identified from the survey and made preliminary decisions on the candidate items. The core team circulated the suggested items to the committee. The steering committee reached the agreements on the relevant items (see *Appendix 6* for details).

Development of the step-by-step selection approach

Guided by the consensus on the selection framework and corresponding items, the core team developed a tentative step-by-step selection approach and tested it using the data of the Western Ontario and McMaster Universities Arthritis Index (WOMAC) obtained from the MID inventory database (up to 2018) [33]. The entire committee then, in a series of meetings, provided feedback to refine the selection approach. When the committee identified concerns and provided suggestions, we tested the revised version with a number of MID estimates in the inventory (e.g., Pain visual analogue scale (VAS), Knee injury and Osteoarthritis Outcome Score (KOOS)) [33]. This iterative process continued through six committee meetings and concluded with the committee members

agreeing on a definitive process of optimal MID selection. During the example-based refinement, the committee reached agreements on issues related to relative and absolute MIDs, and how to deal with the MIDs for the same PROM (or subdomain) that used different scales (see *Appendix 6* for details).

Results

The rationale of the selection approach

As the most important aspect for selecting an anchor-based MID, the committee prioritized the methodological rigor. Thus, we first apply the credibility assessment for all available MID estimates [31] [32] and choose the most credible MIDs (*Figure 5, Step 1; Appendix 7 presents the criteria for choosing the most credible estimates*) for a given PROM (or subdomain). The median of the most credible MIDs for that PROM (or subdomain) constitutes the initial best guess as to an optimal MID [13].

The committee recognized that optimal MIDs may vary by contexts [34]. If such differences exist, the median of the most credible MIDs will not be applicable to all contexts. Evidence suggesting contextual differences might exist include: the most credible MIDs are inconsistent, or the most credible MIDs are consistent with one another but their median is not near the median of all MIDs. Either of these findings requires seeking contextualized factors that can explain the variability among MID estimates *(Figure 5, Step 2)*.

If investigators identify such contextualized factors, they will select the median of the most credible MIDs under each context as the optimal MIDs. If, however, they fail to identify contextualized factors that explain MID variability, they choose the median of the most credible MIDs as the optimal MID.

Explanation of the details of the selection approach

Figure 5 presents the complete processes of the selection approach and *Appendix 7* provides a more detailed narrative description for the selection process. Here follows a summary of the process.

The process begins with identifying the most credible MIDs. To do so, investigators apply the credibility instrument [31], count the number of the five core credibility criteria met, and select the MID estimates with the highest count (*Figure 5, box 1-4; Appendix 7*). The most credible MIDs are those with the highest ratings across the five core credibility criteria (i.e., the highest credibility rank, *Appendix 8*).

Appendix 7 elaborates on how one can assess each credibility criterion and select the most credible MID. Briefly, criterion 1 is rated '*yes or no*'. All other criteria are rated on a 5-point adjectival scale with response options for '*definitely yes' (or 'definitely closely related')* (highest credibility); '*to a great extent'; 'not so much'; 'definitely no' (or 'definitely not related')'* and '*impossible to tell*' (lowest credibility).The best MIDs are those meeting five *'definitely yes/yes/definitely closely related*' for the five core criteria *(Figure 5, box 1).* To identify the most credible MIDs, however, investigators may progressively relax the criteria (*Figure 5, box 4; Appendix 7*) until they find

available MIDs with the highest credibility rank (higher rank means higher credibility, *Appendix 8*). For example, no MIDs for KOOS-quality of life (Qol) met five *'definitely yes/yes/definitely closely related*' (i.e., rank 1). We thus relaxed our criteria and found 2 MIDs with highest credibility rank, referring to them as the most credible MIDs, which were those meeting five '*definitely yes/yes/definitely closely related*' or '*to a great extent*' (i.e., rank 2) (*Appendix 8,9*).

Investigators will then check the consistency of the selected MID estimates (*Figure 5, box 5*) and compare their distribution to the distribution of all available MIDs (*Figure 5, box 9)*. The committee suggested that MIDs be considered consistent with one another if 80% of the estimates lie within an absolute value of 10% of the PROM score above or below their median (i.e., within a range of 20% of the median) *(Box. Glossary*). When the most credible MIDs are consistent (*Figure 5, box 7)*, and the median of the most credible MIDs is near the median of the whole distribution (*Figure 5, box 13*) (the definition for 'near' being the absolute difference between the median of the most credible MIDs and the median of the whole distribution is less than an absolute value of 10% of the PROM score) *(Box. Glossary*), investigators can be confident that the median of the most credible MIDs represents the optimal MID (*Figure 5, box 19; Appendix 7*). For example, we identified the most credible MIDs for the Pain visual analogue scale (VAS-pain) (0-100) were *17, 13.5, 12, 15, 13,16, 14,17* and *20.4*, of which the median was *15*. By definition, they were consistent. Because their median was near the median of the whole distribution (i.e., 15.7)— the absolute difference between them was within 10—we selected the median of the most credible, which was *15*, as the optimal MID (*Appendix 9*).

The most credible estimates may, however, be inconsistent and excessive variability may exist. If that is the case, investigators should try to explain the variability by further consideration of the credibility criteria. The most important credibility criterion addresses the anchor validity: the correlation between the anchor and the PROM of interest [31 34]. Therefore, when the most credible estimates are inconsistent (*Figure 5, box 6*), investigators will prioritize the correlation criterion to explain the remaining variability and among the most credible estimates, select MIDs with high correlations with the anchor ($r \geq 0.5$) [31] (*Figure 5, box 8; Appendix 7*).

The credibility assessment instrument has four additional extension credibility criteria that specifically address the validity of transition rating anchors (*Appendix 7*) [31 35]. After restricting to MIDs with high correlations with the anchor ($r \geq 0.5$) [31], if substantial variability among the estimates remains (*Figure 5, box 17*) - or if no MIDs with high correlations are available (*Figure 5, box 10*), because MIDs were often estimated on transition rating anchors (more than half in the MID inventory) [33], investigators could further consider the additional credibility criteria [31]. We suggest assessing the recall period for the transition rating anchors because except follow-up length, the relevant data for the other three additional criteria were rarely reported [31 36]. Investigators could remove the estimates anchored to transition ratings with a long recall period (>4 weeks) [31] (*Figure 5, box 14*), and determine if the remaining most credible estimates prove consistent (*Figure 5, box 16; Appendix 7*). If, however, no selected estimates are anchored to transition ratings or all the selected estimates anchored to transition ratings are with long recall period (>4 weeks), investigators would skip the assessments (i.e., skip the dotted boxes 14, 16, 20, 21, 23, 26, 27, and 29 in *Figure 5*).

The further consideration of the credibility criteria regarding the anchor validity above, however, may not be necessary. Because there may be few estimates with a high correlation of 0.5 between the PROM and the anchor; it is possible that all the estimates using transition anchors have long recall period; more often, the most credible estimates would appear to be consistent. For example, in our worked examples (*Appendix 9, 10*), we did not further consider the correlation and recall period criteria. After applying the five core credibility assessments, the most credible MIDs were consistent (*Appendix 9, 10*).

At this juncture, investigators will face one of three situations:

i) The (newly) identified most credible MIDs are consistent, and their median is near the median of all MIDs (*Figure 5, box 13, box 25 or box 27; Appendix 7*). If this is the case (referring to the VAS-pain example above), the median of these (newly) identified most credible MIDs represents the optimal MID, applicable to all contexts (*Figure 5, box 19, box 28 or box 29; Appendix 7*).

ii) The (newly) identified most credible MIDs are consistent but their median differs considerably from the median of the whole distribution (*Figure 5, box 12, box 24, box 26; Appendix 7*).

iii) The (newly) identified most credible MIDs are inconsistent (*Figure 5, box 6, box17, or box 20; Appendix 7*).

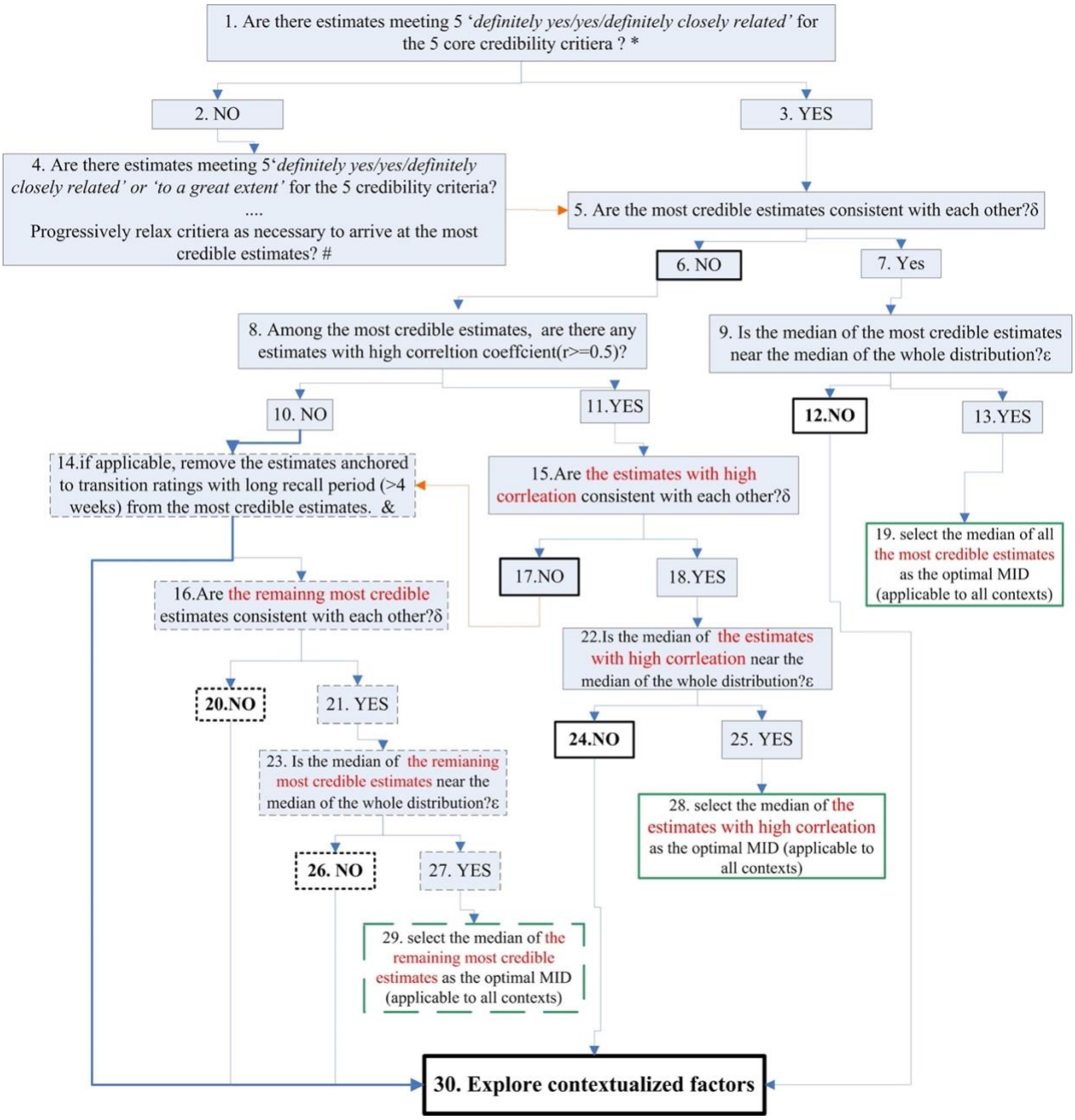
Intuitively, for situation ii), the substantial differences of the two medians may be attributed to MID credibility alone. It is, however, possible that contextualized factors play a role. For example, all the most credible MIDs may be established in a same context (*see the worked example below*),

and the different estimate in another context may be due to the lower credibility or rather due to the different context. Therefore, because either situation ii) or situation iii) suggests the MIDs may be context-dependent, further exploration seeking contextualization as an explanation of variability is required (*Figure 5, box 30; Appendix 7)*.

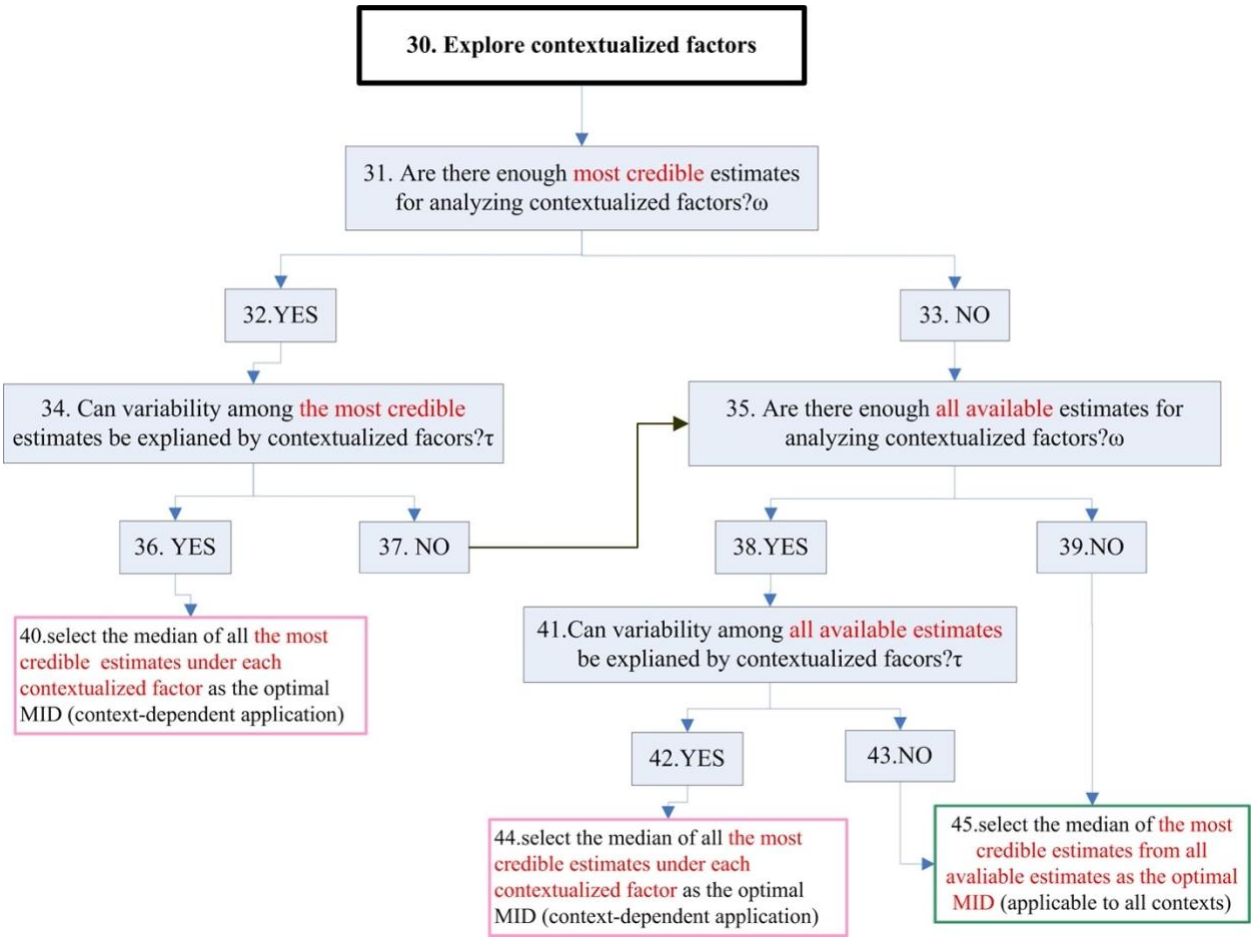
Investigators will use the most credible MID estimates for exploring variability (*Figure 5, box 31*). When, however, the number of credible estimates is insufficient for this exploration (*Figure 5, box 33*) - that is, less than three estimates per contextualized factor (*Box. Glossary*) - investigators will then use all available estimates for the exploration (*Figure 5, box 35*). If the search for contextualized factors yields an explanation for the variability, the optimal MID will be context-dependent: one optimal MID for each context (*Figure 5, box 40 or box 44; Appendix 7*). If no contextual explanation is found, investigators will select the median of the most credible estimates among all available MIDs as the optimal MID. (*Figure 5, box 45; Appendix 7*).

Figure 5. the complete selection process for an optimal anchor-based MD.

Step 1: Choose the most credible MIDs by credibility.



Step 2: Explore contextualized factors.



Footnote:

*Box 1: This is the first step of the selection, where we aim to find the MID estimates with 5 '*definitely yes/yes/definitely closely related*' ratings across the 5 core credibility criteria. The 5 core credibility criteria include: *Q1*. Is the patient or necessary proxy responding directly to both the PROM and the anchor?; *Q2*. Is the anchor easily understandable and relevant for patients or necessary proxy?; *Q3*. Has the anchor shown good correlation with the PROM? (Or, *Q3.1* if the correlation is not reported, to what extent is the construct of the anchor closely related to the construct of the PROM?); *Q4*. Is the MID precise?; *Q5*. Does the threshold or difference between groups on the anchor used to estimate the MID reflect a small but important difference? Response options for *Q1* are *yes or no*. All other criteria are rated on a 5-point adjectival scale with response options for '*definitely yes'* ('*definitely closely related*' for *Q3.1*), '*to a great extent'*, '*not so much'*, '*definitely no'* ('*definitely not related*' for *Q3.1*), and '*impossible to tell'* (*Appendix 7*).

[#]Box 4: The 'most credible estimates' refers to the available MID estimates that receive the highest ratings across the 5 core credibility criteria. To arrive at the most credible estimates, we progressively relax the criteria as follows: first, choose MID estimates with 5 '*definitely yes/yes/definitely closely related'* across the credibility core criteria (Q1, Q2, Q3 (or Q3.1 if correlation coefficient is not reported), Q4 and Q5); if, however, there are no estimates meeting 5 '*definitely yes/yes/definitely closely related*' ratings, we relax our definition of 'most credible' MIDs, and include MIDs that are rated as five '*definitely yes/yes/definitely closely related*' or '*to a great extent*' across the 5 core criteria. If there are no MIDs meeting this relaxed definition, we further relax our criteria to select the estimates in lower credibility rank (see *Appendix 8* for the credibility ranks, higher rank means higher credibility), that is, the estimates with ratings of four '*definitely yes/yes/definitely closely related*' across the core criteria. If not available again, we select MIDs with ratings of four '*definitely yes/yes/definitely closely related*' or '*to a great extent*' across the core criteria and so on (*Appendix 8*). When MIDs with higher rank are available, do not go down the ranking system to select the MIDs with lower rank. For example, if we have one estimate at rank 1, this estimate would be the only one we will use.

[δ] Box 5,15 or 16: consistency is defined as 80% of the MID estimates lying within an absolute value of 10% of the PROM score above or below the median (i.e., within a range of 20% of the median).

[ε] Box 9, 22 or 23: 'near the median of the whole distribution' means that the absolute difference between the median of the selected MID estimates and the median of the whole distribution is less than an absolute value of 10% of the PROM score.

[&] Box 14: only applicable when, among the most credible MID estimates, there are estimates anchored to transition ratings. If, however, no estimates are anchored to transition ratings or all the estimates anchored to transition ratings are with long recall period (>4 weeks), we will skip this step and its downstream steps (i.e., the dotted boxes 14,16, 20, 21, 23, 26, 27 and 29).

[ω] Box 31 or 35: At least 3 most credible estimates per contextualized factor.

[τ] Box 34 or 41: to test whether variability is explained by contextualized factors: compare the distributions; for binary factors, use Wilcoxon Rank Sum Test with a threshold p=0.10.

A worked example for the selection approach

To further illustrate the selection approach, we describe a worked example--the pain subdomain of the WOMAC. In *Appendix 9*, we present more worked examples, including VAS-pain, KOOS-Qol and the 36-item Short Form Survey-mental component summary (SF-36-MCS).

The WOMAC-pain data were obtained from the MID inventory, PROMID (up to 2018) [33]. We described the detailed process for generating the database elsewhere [33]. Briefly, we searched all relevant database to summarize all available anchor-based MIDs of PROMs from primary studies [33]. We identified 13 studies (up to 2018) estimating MIDs for WOMAC-pain. By using different chosen thresholds on the anchor, different anchors, more than one set of participants with different conditions or different analytical anchor-based methods to estimate MID within one study, the 13 studies generated 67 estimates, of which authors expressed 45 MID estimates in absolute terms and 22 MID estimates relative to baseline scores (*Appendix 11*) [33]. Our approach suggests selecting the optimal MIDs from absolute MIDs (*Appendix 6*). We thus conducted the selection for the 45 absolute MIDs. *Appendix 10* presents the entire selection process and *Appendix 12* provides the relevant data for conducting the selection. Below we describe the selection process.

Typically, the WOMAC has five items for pain subdomain assessed on a 0 to 4 scale with a total score ranging from 0 to 20. Although using the same instrument, authors used different scoring system or made conversion of the scores and thus expressed the results with different scales (scales ranging from 0 to 10, 0 to 20, 0 to 50 and 0 to 100, *Appendix 11*). Therefore, we transformed each estimate into a scale of 0 to 100. Following this transformation, the MIDs ranged from 0 to 35 (*Appendix 11*).

After assessing MID credibility (*Figure 5, box 1*), no estimates met five '*definitely yes/yes/definitely closely related*' ratings across the five core credibility criteria (*Appendix 10, 12*). By relaxing the criteria *(Figure 5, box 4);* we found 5 estimates meeting five *'definitely yes/yes/definitely closely related*' or '*to a great extent'* (*Appendix 10,12)*. *Table 4* presents the absolute value of these most credible MID estimates. Using our definition for consistency *(Box. Glossary)*, these most credible MID estimates were consistent (*Figure 5, box 7; Appendix 10*). Because the median of the most credible estimates (*28.1)* was not near the median of the whole distribution (*12.5*) (*Box. Glossary; Figure 5, box 12; Appendix 10*), we postulated that the MIDs for WOMAC-pain could be context-dependent. We therefore explored the possibility that contextualized factors could explain MID variability.

Because the number of most credible MID estimates, five, was not sufficient for the exploration *(Figure 5, box 33, Appendix 10)*, we instead used all available estimates (*Figure 5, box 35; Appendix 10*). We explored the impact of patient condition (knee complaints vs hip complaints) on the variability of the MID estimates by comparing the medians of the two groups using Wilcoxon Rank Sum Test. We found the patient condition did not explain the variability. We then explored whether the intervention (surgical vs non-surgical intervention) explained the variability and found a significant difference between MIDs generated in surgical versus non-surgical intervention settings (*p=0.009, Table 4; Appendix 10,12*).

Therefore, the selection and application of the optimal MID for WOMAC-pain was context-dependent—the optimal MID differed depending on whether the patients were undergoing surgical

intervention (*Figure 5, box 44; Appendix 10*). That is, there proved to be two optimal MIDs: one selected from the MIDs estimated under the context of surgical intervention that should be exclusively used in the context of surgical intervention; and the other selected from the MIDs estimated under the context of non-surgical intervention that should be exclusively used in the context of non-surgical intervention.

Under the context of surgical intervention, we found the most credible estimates were those meeting 5 '*definitely yes/yes/definitely closely related'* or '*to a great extent'* across the five core credibility criteria, including *29.26, 29.9, 20.5, 28.1, 23.5*, with a median of *28.1 (Table 4*). The most credible estimates for non-surgical intervention were those meeting 4 '*definitely yes/yes/definitely closely related*' or '*to a great extent'* across the five core credibility criteria, including *11.8, 12.9, 6.4, 8.3, 13.51, 8.74, 15, 8.74, 7.09, 4.1,* with a median of *8.7 (Table 4).* Correspondingly, the optimal MID for surgical intervention was *28.1* applicable to the context of surgical intervention, and the optimal MID for non-surgical intervention was *8.7* applicable to the context of non-surgical intervention.

Table 4. The optimal anchor-based MIDs expressed in absolute terms for WOMAC-pain (up to 2018).

	N	Point estimates[#]	Median[#]
All MIDs	45	-	12.0[&]
'Most credible' MIDs	5	29.26, 29.9, 20.5, 28.1, 23.5.	28.1
All MIDs under the context of surgical intervention	17	-	23.5
'Most credible' MIDs	5	29.26, 29.9, 20.5, 28.1, 23.5.	28.1
The optimal MID	-	**28.1**	-
All MIDs under the context of non-surgical intervention	28		9.0
'Most credible' MIDs	10	11.8, 12.9, 6.4, 8.3[*], 13.51, 8.74, 15, 8.74, 7.09, 4.1	8.7
The optimal MID	-	**8.7**	-

Footnote:

See *Appendix 11* for all MID estimates. The most credible estimates for all MIDs and MIDs under the context of surgical intervention were those meeting 5 '*definitely yes/yes/definitely closely related*' or '*to a great extent*' across the five core credibility criteria, while the most credible estimates under the context of non-surgical intervention were those meeting 4 '*definitely yes/yes/definitely closely related*' or '*to a great extent'* across the five core credibility criteria.

[#]We used the absolute value of the estimates for calculating the median. All the estimates were transformed into a 0 to 100 scale.

[&] There was an estimate (14.66) that the authors did not report the lower and upper value of the scale. We excluded it and thus a total of 44 available estimates was used to for calculation the median. If add the excluded estimate (14.66) and use all the 45 estimates, the median was 12.1.

*This estimate was -0.83 estimated on a 0-10 scale. We used the absolute value and transformed it into a 0 to 100 scale and got 8.3.

Discussion

Main findings

Based on expert experience, a systematic survey, and example-based refinement, we developed the first systematic step-by-step approach for selecting an optimal anchor-based MID from various MID estimates of a given PROM. We have successfully applied the approach to several PROMs in the MID inventory [33] (*Appendix 9,10)*.

This approach bases the selection on explaining the variability of all available MID estimates for the PROM of interest. We prioritize the methodological rigor of MID estimation and, through credibility assessment [31], select the most credible MID estimates (*Figure 5, Step 1*). If the most credible MIDs fall in a relatively narrow range, investigators choose the median as the optimal MID.

If, however, the most credible MIDs are consistent but their median differs substantially from the median of all the MIDs, or the most credible MIDs are inconsistent, credibility alone cannot explain the variation and there may exist contextualized factors influencing MID estimates [34]. Our approach mandates further exploring contextualized factors to explain the variability among all the MIDs (*Figure 5, Step 2*). The potential contextualized factors that deserve a consideration could come from the suggestions of previous researchers and include intervention (e.g., surgical vs conservative treatments), patient condition (e.g., knee vs hip osteoarthritis), baseline disease severity, patient age, follow-up duration, socioeconomic status, geography, and sex [34].

When the process identifies contextualized factors, the median of the most credible MIDs under a specific context represents the optimal MID and we arrive at context-dependent MIDs. If the process fails to identify contextualized factors that explain MID estimate variability, investigators will still select the median of MIDs with highest credibility as the optimal MID and apply it to all contexts.

Strength and limitations

Strengths of the study include the range and depth of expertise of the study team, the systematic survey that informed the process [34], iterative modification of the selection approach based on expert feedback and the application of the approach to PROMs in our MID inventory [33], and the resultant transparent workable process. The approach worked well in selecting optimal MIDs for common PROMs in the inventory [33] (*Appendix 9,10*).

Our study has limitations. The selection process is complex and may be burdensome. Before navigating the selection process, users must collect all available MID estimates for the PROM of interest. Our MID inventory online platform (www.promid.org), can however, provide the necessary material.

Suboptimal reporting of MID estimation studies (e.g., the lack of the upper and lower limit of the PROM scale and all the relevant information about MID credibility) may lead to the difficulties in the selection. Because the data for exploring contextualized factors may be limited, there could be some factors that are important but not measured or reported by authors that in fact are responsible for differences between MIDs. If such unmeasured or unreported differences exist and are substantial, they would limit the applicability of the optimal MID selected by our approach. The

selection may not work well when only a few yet divergent MID estimates are available. Having established an optimal MID, when new estimates emerge, investigators may need to review the process. The selection approach includes thresholds that are somewhat arbitrary. For instance, we considered a difference less than an absolute value of 10% score of the PROM scale as a relatively narrow range (i.e., the definition of 'consistency' and 'near the median', see *Box. Glossary*). Our steering committee, a small group of experts, may not be representative. The committee members, however, were diverse in geography and sex, and a number (GHG, MTK, CBT, TAF, BT) have enormous experience in health measurement and have been working with the health measurement research community for decades. Finally, because no consensus on the gold standard of 'optimal MID', exists, future insights on anchor-based MID may require the modification of the selection approach.

Implications

In recent years, PROMs have become increasingly popular in clinical practice and clinical trials [37-42]. PROMs provide crucial information regarding treatment efficacy from patients' perspectives that cannot be captured by other outcomes. Along with the use of PROMs, the number of anchor-based MID estimates, as well as the demand for a suitable MID to aid the interpretation of treatment effects, has increased considerably [20-22 41]. Failure to identify an optimal MID might result in serious misinterpretations of PROM results. Further, available MID estimates often vary widely [42-44], presenting a dilemma for those conducting clinical trials, authors of systematic reviews, guideline developers, clinicians, funders, and policy makers: *how to choose the best MID?* The selection approach described here addresses that issue by providing a logical and systematic way to select an optimal MID for a PROM when multiple discrepant MIDs exist.

The selection process is geared to explaining the variability among available MID estimates and where appropriate, to provide a single optimal MID, i.e., the estimate taking the median of the selected estimates in a relatively narrow range [13]. Two parts, the methodological rigor, and the generalizability (i.e., factors influencing the MID application), frame the selection process. The selection covers the important issues about estimation and application of anchor-based MIDs, however, does not address the analytical methods used to estimate anchor-based MID (e.g., mean change method, ROC method) [34].

We could have used a formal Delphi process to choose candidate items to inform the first draft of the selection approach that subsequently underwent iterative example-based refinement. Indeed, it might have been preferable to do so. The choice of candidate items was, however, straightforward, and it is unlikely that appreciable differences would have emerged from the Delphi panel (see *Appendix 6, eTable 1*).

Typically, if an intervention has more associated burdens and adverse effects, people will require a larger effect or improvement and thus a larger MID. This scenario is what the data of WOMAC-pain demonstrated: the MIDs for surgery were larger than non-surgery. Thus, to use responder analysis for analyzing the benefits of the interventions, researchers should dichotomize the participants using the MIDs specific to the intervention they received. When researchers use an MID in the process to choose a target difference to calculate sample size [45] [46] and interpret results measured by mean difference for clinical trials [46], they should take the interventions into account. For example, for a trial measuring the effects of surgery vs non-surgery on WOMAC-pain, because the treatment difference between the interventions would reflect the demand of an improvement

for surgery, using the optimal MID for surgical intervention would be more appropriate, and the sample size calculated accordingly.

In our worked examples, only a small proportion of MIDs proved of high credibility (e.g., WOMAC-pain had 5 highly credible estimates out of 45 published estimates). This finding highlights the need for more high-quality studies to establish new credible MID estimates and better reporting of MID studies [36]. The criteria in our credibility tool [31] provide key methodological rigor for developing trustworthy anchor-based MIDs.

Ideally, after selection, the optimal MID would not change as new evidence emerges and thus become a unique standard to aid the interpretation for a given PROM. When that occurs, generating new estimates for such a PROM would be a poor use of limited research resources. The larger the number of high credibility, consistent MIDs, the more compelling the case that a definitive optimal MID has been established. What threshold one should use for this conclusion remains, however, open to debate and therefore another potential area for subsequent research.

Acknowledgement

The authors thank Mark Phillips, Bradley C Johnston, Dena Zeraatkar, Meha Bhatt, Xuejing Jin, Romina Brignardello-Petersen, Olivia Urquhart, Farid Foroutan, Stefan Schandelmaier, Hector Pardo-Hernandez, Robin WM Vernooij, Hsiaomin Huang, Linan Zeng, Yamna Rizwan, Reed Siemieniuk, Lyubov Lytvyn, Zhikang Ye, Liam Yao, Vanessa Wong, Donald L Patrick, Shanil Ebrahim, Gihad Nesrallah, Holger J Schunemann, Mohit Bhandari, and Lehana Thabane for their contributions on the MID inventory and credibility instrument projects. The authors also thank Anila Qasim for her contributions to maintain the MID inventory online platform (www.promid.org).

Footnotes

Contributors

GHG, TD and YW initiated the project. AC-L, BT, CBT, GHG, MTK, MW, TD, TAF and YW provided insights on the selection framework and contributed to the consensus of the items used to develop the selection approach. AC-L, TD, YW and GHG drafted the details of the selection approach. All authors provided feedback for the revision of the selection approach. YW was responsible for revising the selection approach until all authors reached agreements. YW wrote the initial draft of the manuscript and all other authors reviewed and revised the manuscript draft. All authors approved the final version of the manuscript. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding

No specific funding was given to this study.

Competing Interests

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: no support for the submitted work. AC-L, GHG and TD have a patent issued for the Credibility instrument for judging the minimal important difference and the Patient Reported Outcome Minimal Important Difference (PROMID) Database (www.promid.org). AC-L, CBT, GHG, MW, TAF, TD reports grants, pending patents, personal fees, roles in advisory board or leadership in the committees outside the submitted work; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval

The ethical approval was not required for this study.

Patient consent

The patient consent was not required for this study.

Data sharing

The raw data of the worked example are available on PROMID database (http://www.promid.org).

Provenance and peer review

Not commissioned, externally peer reviewed.

Patient and public involvement

Patients and the public were not involved in the design, conduct, or reporting in this methodological research, as our selection approach 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.

Copyright statement

The Credibility instrument for judging the trustworthiness of minimal important difference estimates, authored by Dr Devji et al., is the copyright of McMaster university (copyright 2018, McMaster university). The Minimal Important Difference Inventory, authored by Dr Carrasco-Labra et al., is the copyright of McMaster University (Copyright 2018, McMaster University, Hamilton, Ontario, Canada). The Credibility instrument and The Minimal Important Difference Inventory have been provided under license from McMaster University and must not be copied, distributed, or used in any way without the prior written consent of McMaster University. Contact the McMaster Industry Liaison Office at McMaster University, email: milo@mcmaster.ca for licensing details.

References

1. Kluzek S, Dean B, Wartolowska KA. Patient-reported outcome measures (PROMs) as proof of treatment efficacy. *BMJ Evid Based Med* 2022;27(3):153-55. doi: 10.1136/bmjebm-2020-111573 pmid: 34088713
2. Valderas JM, Kotzeva A, Espallargues M, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual Life Res* 2008;17(2):179-93. doi: 10.1007/s11136-007-9295-0 pmid: 18175207
3. Basch E, Deal AM, Kris MG, et al. Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial. *J Clin Oncol* 2016;34(6):557-65. doi: 10.1200/JCO.2015.63.0830 pmid: 26644527
4. European Medicines Agency. Committee for Medicinal Products for Human Use (CHMP): Appendix 2 to the Guideline on the evaluation of anticancer medicinal products in man. The use of patient-reported outcome (pro) measures in oncology studies. 2016. [Accessed Sep 26, 2022]. Available from: https://www.ema.europa.eu/en/documents/other/appendix-2-guideline-evaluation-anticancer-medicinal-products-man_en.pdf
5. Food and Drug Administration. Guidance for Industry: patient-reported outcome measures: use in medical product development to support labelling claims. 2009. [Accessed June 28, 2022]. Available from: https://www.fda.gov/media/77832/download
6. Calvert M, Blazeby J, Altman DG, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *Jama* 2013;309(8):814-22. doi: 10.1001/jama.2013.879 pmid: 23443445
7. Coens C, Pe M, Dueck AC, et al. International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium. *Lancet Oncol* 2020;21(2):e83-e96. doi: 10.1016/s1470-2045(19)30790-9 pmid: 32007209
8. Calvert M, Kyte D, Mercieca-Bebber R, et al. Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension. *Jama* 2018;319(5):483-94. doi: 10.1001/jama.2017.21903 pmid: 29411037
9. Brignardello-Petersen R, Carrasco-Labra A, Shah P, et al. A practitioner's guide to developing critical appraisal skills: what is the difference between clinical and statistical significance? *J Am Dent Assoc* 2013;144(7):780-6. doi: 10.14219/jada.archive.2013.0187 pmid: 23813258
10. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10(4):407-15. doi: 10.1016/0197-2456(89)90005-6 pmid: 2691207
11. Schünemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res* 2005;40(2):593-7. doi: 10.1111/j.1475-6773.2005.00374.x
12. 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 pmid: 20937092
13. Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77(4):371-83. doi: 10.4065/77.4.371 pmid: 11936935
14. Turner D, Schünemann HJ, Griffith LE, et al. The minimal detectable change cannot reliably replace the minimal important difference. *J Clin Epidemiol* 2010;63(1):28-36. doi: 10.1016/j.jclinepi.2009.01.024 pmid: 19800198
15. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *Jama* 2014;312(13):1342-3. doi: 10.1001/jama.2014.13128 pmid: 25268441
16. Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life-a systematic review. *J Clin Epidemiol* 2017;89:188-98. doi:

10.1016/j.jclinepi.2017.06.009 pmid: 28676426
17. Terwee CB, Peipert JD, Chapman R, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. *Qual Life Res* 2021;30(10):2729-54. doi: 10.1007/s11136-021-02925-y pmid: 34247326
18. 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. doi: 10.1136/bmjopen-2018-028777 pmid: 30787096
19. 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. doi: 10.1136/bmjopen-2016-015587 pmid: 28495818
20. Okereke OI, Reynolds CF, 3rd, Mischoulon D, et al. Effect of Long-term Vitamin D3 Supplementation vs Placebo on Risk of Depression or Clinically Relevant Depressive Symptoms and on Change in Mood Scores: A Randomized Clinical Trial. *Jama* 2020;324(5):471-80. doi: 10.1001/jama.2020.10224 pmid: 32749491
21. Fiore JF, Jr., El-Kefraoui C, Chay MA, et al. Opioid versus opioid-free analgesia after surgical discharge: a systematic review and meta-analysis of randomised trials. *Lancet* 2022;399(10343):2280-93. doi: 10.1016/s0140-6736(22)00582-7 pmid: 35717988
22. Chahal J, Whelan DB, Hoit G, et al. Anterior Cruciate Ligament Patellar Tendon Autograft Fixation at 0° Versus 30° Results in Improved Activity Scores and a Greater Proportion of Patients Achieving the Minimal Clinical Important Difference For Knee Injury and Osteoarthritis Outcome Score Pain: A Randomized Controlled Trial. *Arthroscopy* 2022;38(6):1969-77. doi: 10.1016/j.arthro.2021.12.018 pmid: 34952186
23. Koynova D, Lühmann R, Fischer R. A Framework for Managing the Minimal Clinically Important Difference in Clinical Trials. *Ther Innov Regul Sci* 2013;47(4):447-54. doi: 10.1177/2168479013487541 pmid: 30235520
24. Draak THP, de Greef BTA, Faber CG, et al. The minimum clinically important difference: which direction to take. *Eur J Neurol* 2019;26(6):850-55. doi: 10.1111/ene.13941 pmid: 30793428
25. Engel L, Beaton DE, Touma Z. Minimal Clinically Important Difference: A Review of Outcome Measure Score Interpretation. *Rheum Dis Clin North Am* 2018;44(2):177-88. doi: 10.1016/j.rdc.2018.01.011 pmid: 29622290
26. 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(7):725-8. doi: 10.1016/j.jclinepi.2008.09.012 pmid: 19128938
27. Revicki DA, Cella D, Hays RD, et al. Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes* 2006;4:70. doi: 10.1186/1477-7525-4-70 pmid: 17005038
28. Malec JF, Ketchum JM. A Standard Method for Determining the Minimal Clinically Important Difference for Rehabilitation Measures. *Arch Phys Med Rehabil* 2020;101(6):1090-94. doi: 10.1016/j.apmr.2019.12.008 pmid: 31953077
29. Wright A, Hannon J, Hegedus EJ, et al. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). *J Man Manip Ther* 2012;20(3):160-6. doi: 10.1179/2042618612y.0000000001 pmid: 23904756
30. Leidy NK, Wyrwich KW. Bridging the gap: using triangulation methodology to estimate minimal clinically important differences (MCIDs). *Copd* 2005;2(1):157-65. doi: 10.1081/copd-200050508 pmid: 17136977
31. 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. doi: 10.1136/bmj.m1714 pmid: 32499297
32. Wang Y, Devji T, Carrasco-Labra A, et al. An extension MID credibility item addressing construct proximity is a reliable alternative to correlation item. *J Clin Epidemiol* 2023 doi: 10.1016/j.jclinepi.2023.03.001 pmid: 36878330

33. 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. doi: 10.1016/j.jclinepi.2020.11.024 pmid: 33321175
34. Wang Y, Devji T, Qasim A, et al. A systematic survey identified methodological issues in studies estimating anchor-based minimal important differences in patient-reported outcomes. *J Clin Epidemiol* 2022;142:144-51. doi: 10.1016/j.jclinepi.2021.10.028 pmid: 34752937
35. Guyatt GH, Norman GR, Juniper EF, et al. A critical look at transition ratings. *J Clin Epidemiol* 2002;55(9):900-8. doi: 10.1016/s0895-4356(02)00435-3 pmid: 12393078
36. Carrasco-Labra A, Devji T, Qasim A, et al. Serious reporting deficiencies exist in minimal important difference studies: current state and suggestions for improvement. *J Clin Epidemiol* 2022;150:25-32. doi: 10.1016/j.jclinepi.2022.06.010 pmid: 35760237
37. Vodicka E, Kim K, Devine EB, et al. Inclusion of patient-reported outcome measures in registered clinical trials: Evidence from ClinicalTrials.gov (2007-2013). *Contemp Clin Trials* 2015;43:1-9. doi: 10.1016/j.cct.2015.04.004 pmid: 25896116
38. Snyder CF, Aaronson NK, Choucair AK, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. *Qual Life Res* 2012;21(8):1305-14. doi: 10.1007/s11136-011-0054-x pmid: 22048932
39. Black N. Patient reported outcome measures could help transform healthcare. *Bmj* 2013;346:f167. doi: 10.1136/bmj.f167 pmid: 23358487
40. Mercieca-Bebber R, Williams D, Tait MA, et al. Trials with patient-reported outcomes registered on the Australian New Zealand Clinical Trials Registry (ANZCTR). *Qual Life Res* 2018;27(10):2581-91. doi: 10.1007/s11136-018-1921-5 pmid: 29915979
41. Liu KY, Schneider LS, Howard R. The need to show minimum clinically important differences in Alzheimer's disease trials. *Lancet Psychiatry* 2021;8(11):1013-16. doi: 10.1016/s2215-0366(21)00197-8 pmid: 34087114
42. Keurentjes JC, Van Tol FR, Fiocco M, et al. Minimal clinically important differences in health-related quality of life after total hip or knee replacement: A systematic review. *Bone Joint Res* 2012;1(5):71-7. doi: 10.1302/2046-3758.15.2000065 pmid: 23610674
43. Riepe C, Osada N, Reich A, et al. Minimal Clinically Important Difference in Chronic Pruritus Appears to be Dependent on Baseline Itch Severity. *Acta Derm Venereol* 2019;99(13):1288-90. doi: 10.2340/00015555-3332 pmid: 31580464
44. Rodrigues JN, Mabvuure NT, Nikkhah D, et al. Minimal important changes and differences in elective hand surgery. *J Hand Surg Eur Vol* 2015;40(9):900-12. doi: 10.1177/1753193414553908 pmid: 25320122
45. Cook JA, Julious SA, Sones W, et al. DELTA(2) guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *Bmj* 2018;363:k3750. doi: 10.1136/bmj.k3750 pmid: 30560792
46. Brozek 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 pmid: 17005037

Appendix 6. Additional methods for developing the selection approach.

Decisions on the candidate items

After we gathered the committee's view on the selection framework—MID methodological rigor and generalizability, we reached consensus on items identified from the survey [1] to be included in the selection framework (eTable 1), which informed the following development of the systematic selection approach. A core team (AC-L, GHG, TD and YW) first conducted intensive discussions on the candidate items identified from the survey and made preliminary decisions on the candidate items. The core team circulated the suggested items to the committee. The steering committee reached the following agreements.

1. MID methodological rigor

The literature review revealed that the existing anchor-based MID credibility assessment tool captures the key issues for a credible MID estimate [1 2]. These include, the five core criteria (patient-rated anchor; interpretable and relevant anchor; precise MID estimate; good correlation between the anchor and the PROM (at least 0.5); and a threshold on the anchor reflecting a small but important difference) and the four additional criteria for transition rating anchors (the optimal follow-up length; satisfactory correlation between transition item and PROM score at follow-up; the correlation between transition item and PROM score at baseline being equal but opposite to the correlation between transition item and PROM score at follow-up; and appreciably greater correlation of the transition item with PROM change score than the correlation of the transition item with PROM score at follow-up) (*Appendix 7*) [2].

The committee identified one limitation: if the correlation between anchor and PROM is unavailable, assessment must include the assessment of construct proximity between the anchor

and the PROM to substitute for the correlation criterion [1]. That is, MIDs estimated from very closely related constructs of the anchor and PROM (e.g., both measure pain) would likely result in a stronger correlation than those estimated from anchors with constructs that differ from the PROM (e.g., PROM measures pain vs anchor measures how worthwhile is the surgery) (*Appendix 7*). Our parallel work has addressed this issue [3].

2. MID generalizability

2.1. Contextualized factors

Investigators could identify contextualize factors from the MID estimation studies and explore the impacts of these contextualized factors on MID variability. When evidence suggests contextualized factors could explain the variability of the MID estimates, to ensure the MID applicable to the contexts of interest, the committee agreed that selecting and applying the MID must be context-dependent [1]. According to the survey [1], patient and intervention characteristics, could be the possible contextualized factors for exploration.

2.2 MIDs for improvement and deterioration

The survey identified that several authorities advocate for distinguishing MID for improvement and deterioration[1]. However, authors of papers reporting the development of MIDs often do not consider the direction; instead, they estimate the absolute value of change scores [4-6]. If compelling evidence of a difference is available, separate MIDs for improvement and deterioration are necessary. If there is no evidence for a difference, the committee suggests not distinguishing between the MIDs for improvement and deterioration, and thus using the absolute value of the two MIDs interchangeably.

The additional issues addressed during the example-based refinement

When developing the step-by-step selection approach based on the real data from the MID inventory [7], we iteratively refined our selection approach. During the example-based refinement, the committee reached agreements on issues related to relative and absolute MIDs, and how to deal with the MIDs for the same PROM (or subdomain) that used different scales.

1. Relative and absolute MIDs

There are two ways to establish an anchor-based MID: using absolute measures of change or using relative measures of change (i.e., change as a percentage of baseline score) [8]. When investigators established MIDs using both the absolute and relative approaches, an optimal MID for each is possible. Because of the likely advantages of the absolute approach over the relative [8 9], the committee suggested using, if available, the optimal MIDs established from absolute approach.

2. The same PROM (or subdomain) in which authors use different scoring approaches

PROMs are typically designed to be treated as continuous outcomes. For a same PROM or subdomain of a PROM, investigators may use different scoring algorithms and thus report scores with different scale ranges (e.g., the pain subdomain of the WOMAC has scale range of 0 to 50 and 0 to 100). When this is the case, the committee suggested, before the selection, transforming all the estimates into a same scale.

eTable 1. The list of identified Items relevant to selecting anchor-based MID*

Code	29 candidate Items relevant to MID selection identified from the survey	Item included for the development of the selection approach (yes/ no)	Reasons for inclusion/exclusion
Anchor perspective	The change being perceived as minimally important on the anchor would be perceived as beneficial from the patient's viewpoint.	Yes	Under the MID selection framework of MID methodology: an MID credibility criterion (Q1#)
	Proxy perspective is acceptable if patients cannot make decisions.	Yes	Under the MID selection framework of MID methodology: an MID credibility criterion (Q1#)
	Estimating MID should use clinical evaluation as anchor.	Yes	Under the MID selection framework of MID methodology: an MID credibility criterion (Q1#)
Anchor threshold	The choice of what constitutes an MID on the anchor should reflect a small but important difference .	Yes	Under the MID selection framework of MID methodology: an MID credibility criterion (Q5#)
Anchor interpretability	Anchor should be easily understandable.	Yes	Under the MID selection framework of MID methodology: an MID credibility criterion (Q2#)
Correlation between PROM and anchor	At least moderate correlation between PROM and anchor.	Yes	Under the MID selection framework of MID methodology: an MID credibility criterion (Q3#)
	Transition scale should correlate equally strongly with the pretest and posttest PROM scores but with opposite signs.	Yes	Under the MID selection framework of MID methodology: additional MID credibility criteria for transition anchors#
	The correlation of transition scale with change scores on PROM should be at least 0.2 greater (in absolute terms) than the correlations with either the baseline or the follow-up test scores.	Yes	Under the MID selection framework of MID methodology: additional MID credibility criteria for transition anchors#
Anchor construct	Anchor and PROMs should measure the same or closely related constructs.	Yes	Under the MID selection framework of MID methodology: an MID credibility criterion (Q3.1#)
Recall time of transition scale	The recall time should not be too long when using transition scale as anchor.	Yes	Under the MID selection framework of MID methodology: additional MID credibility criteria for transition anchors#

Unchanged group of transition scale	The use of unchanged group when using transition scale as anchor.	NO	The additional MID credibility criteria for transition anchors have captured the assessments of the validity of transition anchors.
Number of anchors	Multiple anchors should be used to measure MID, rather than single anchor.	NO	Few anchor-based MIDs were established from multiple anchors.
Measurement properties of anchor	Satisfactory anchor measurement properties (validity and reliability) are the prerequisite of an optimal anchor.	NO	Reliability is rarely reported; the MID credibility criteria have captured the assessment of anchor validity.
Measurement properties of PROM	Required measurement properties for PROM when measuring MID: reliability, validity and responsiveness.	NO	We select optimal MIDs from existing MIDs assuming that all the corresponding PROMs were well established.
	Ordinal scales do not support the mathematical calculations required for MCID calculation.	NO	The selection approach will choose optimal MIDs from exiting MIDs, assuming that to calculate the MIDs, the corresponding PROMs have interval scales.
MID varying by contexts	MID may vary depending on patient characteristics.	Yes	Under the MID selection framework of MID application contexts
	MID may vary depending on follow-up length.	Yes	Under the MID selection framework of MID application contexts
	MID may vary depending on intervention.	Yes	Under the MID selection framework of MID application contexts
MID direction	Distinguishing MID for improvement vs deterioration.	Yes	Under the MID selection framework of MID application contexts
Analytical method	MID may vary depending on the analytical method used.	NO	The main analytical methods used to estimate MID include mean difference, mean change and ROC approaches. It remains uncertain whether ROC approaches are superior or inferior to mean change and/or mean difference methods. Thus, analytical methods cannot yet stand as a selection criterion.
Generic instruments	For generic instruments, MID should not vary depending on the contexts.	NO	The MID selection framework considers the impact of the contextualized factors.

			Whether the impact of contextualized factors matters for generic instruments will be decided by the evidence.
Precision of MID	Precision of MID	Yes	Under the MID selection framework of MID methodology: an MID credibility criterion (Q4[#])
MID in relation to measurement error	MID should be beyond measurement error.	NO	For anchor-based MIDs, authors rarely reported the associated measurement error, making it difficult to decide the relationship between MID and measurement error. Thus, this cannot yet stand as a credibility criterion.
Triangulation	Triangulation of distribution-based and anchor-based methods.	NO	We select optimal MIDs from anchor-based MIDs.
Representative sample	Representative sample	NO	No detailed guidance on what constitute a representative sample from which an optimal MIDs should be calculated.
MID selection	MID selection: match study context	Yes	Under the MID selection framework of MID application contexts
	MID selection: evidence-based and consensus processes	NO	No detailed guidance on the processes.
	MID selection: MID levels	NO	No clear guidance on the MID levels; the MID credibility criteria distinguish trustworthy and untrustworthy MIDs.
A single MID or range of plausible MIDs	A single value or range of plausible MIDs is appropriate for many PROMs	Yes	The selection considers the median of the most credible MID.

*This table was reproduced from the Table 2 of the systematic survey [1].

See Appendix 7 for all the MID credibility criteria.

Reference

1. Wang Y, Devji T, Qasim A, et al. A systematic survey identified methodological issues in studies estimating anchor-based minimal important differences in patient-reported outcomes. *J Clin Epidemiol* 2022;142:144-51. doi: 10.1016/j.jclinepi.2021.10.028
2. 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. doi: 10.1136/bmj.m1714
3. Wang Y, Devji T, Carrasco-Labra A, et al. An extension MID credibility item addressing construct proximity is a reliable alternative to correlation item. *J Clin Epidemiol* 2023 doi: 10.1016/j.jclinepi.2023.03.001
4. Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. *Phys Ther* 2006;86(5):735-43.
5. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res* 2011;11(2):171-84. doi: 10.1586/erp.11.9
6. Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life-a systematic review. *J Clin Epidemiol* 2017;89:188-98. doi: 10.1016/j.jclinepi.2017.06.009
7. 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. doi: 10.1016/j.jclinepi.2020.11.024
8. Zhang Y, Zhang S, Thabane L, et al. Although not consistently superior, the absolute approach to framing the minimally important difference has advantages over the relative approach. *J Clin Epidemiol* 2015;68(8):888-94. doi: 10.1016/j.jclinepi.2015.02.017
9. Terluin B, Roos EM, Terwee CB, et al. Assessing baseline dependency of anchor-based minimal important change (MIC): don't stratify on the baseline score! *Qual Life Res* 2021;30(10):2773-82. doi: 10.1007/s11136-021-02886-2

Appendix 7. Detailed narrative description for the selection approach.

Summary

This approach is geared to explaining the variability of all available anchor-based MID estimates for a given PROM by the methodological rigor and contextualized factors influencing the MID application, and where appropriate, provides a single optimal MID, i.e., the median of the selected estimates in a relatively narrow range.

We prioritize the methodological rigor of MID estimation and, through credibility assessment, select the most credible MID estimates (*Figure 5, Step 1*). If the most credible MIDs fall in a relatively narrow range, we select the median as the optimal MID.

If, however, the most credible MIDs are consistent but their median differs substantially from the median of all the available MIDs, or the most credible MIDs are inconsistent, credibility alone cannot explain the variation and there may exist contextualized factors influencing MID estimates. This approach then mandates further exploring contextualized factors, including patient and intervention characteristics, to explain the variability among the MIDs (*Figure 5, Step 2*).

Before the selection

There are two ways to establish an anchor-based MID: using absolute measures of change or using relative measures of change (i.e., change as a percentage of baseline score) [1]. When MIDs for a given PROM established from both the absolute and relative approaches, our selection approach suggests using the optimal MIDs established from the absolute approach.

For a same PROM or subdomain of a PROM, investigators may use different scoring algorithms and thus report scores with different scale ranges (e.g., the pain subdomain of the WOMAC has scale range of 0 to 50 and 0 to 100). When this is the case, our approach suggests, before the selection, transforming all the MID estimates into a same scale.

The users could formulate a prior hypothesis of the important contextualized factors and abstract the factors from the MID estimation studies. When evidence suggests contextual differences might exist among the MIDs, users can further explore the impact of these contextualized factors on MIDs. The potential contextualized factors that deserve a consideration could come from the suggestions of previous researchers, including intervention (e.g., surgical vs conservative treatments), patient condition (e.g., knee vs hip osteoarthritis), baseline disease severity, patient age (e.g., young vs old), follow-up length, socioeconomic status, geography, and sex [2]. It is likely, however, so far, the data to explore the factors are limited. Users should take advantage of all the data that are available for the factors and acknowledge that there may be other factors that could impact the MID variability, but the data are not available.

Detailed narrative description of the selection approach

Step 1. Choose the most credible MIDs by credibility.

A. MID credibility assessment

We use a validated instrument to evaluate the credibility of anchor-based MIDs for PROMs [3 4]. The instrument provides a systematic approach to deal with the methodological assessment for anchor-based MIDs. The instrument includes two components: 1) the five core criteria applicable to any anchor-based MID estimation and 2) four extension criteria addressing transition rating – also referred to as global ratings of change – anchors.

1. Core criteria

Q1. Is the patient or necessary proxy responding directly to both the PROM and the anchor?

Elaboration: *Q1 is rated 'yes or no'. 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) require a proxy respondent. We suggest both patient or necessary proxy responses are credible.*

Q2. Is the anchor easily understandable and relevant for patients or necessary proxy?

Elaboration: *Q2 is rated on a 5-point adjectival scale with response options for 'definitely yes' (highest credibility); 'to a great extent'; 'not so much'; 'definitely no' and 'impossible to tell' (lowest credibility). 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, status on an important and easily understood measure of function, the presence of symptoms, disease severity, response to treatment, or the prognosis for future events, such as death, use of healthcare facilities, or job loss.*

Q3. Has the anchor shown good correlation with the PROM?

Elaboration: *Q3 is rated on a 5-point adjectival scale with response options for 'definitely yes' (r>0.7; highest credibility); 'to a great extent' (r ≥ 0.5 to r <0.7); 'not so much' (r ≥ 0.3 to r <0.5); 'definitely no' (r <0.3) and 'impossible to tell' (lowest credibility).*

Q3.1 if the correlation is not reported, to what extent is the construct of the anchor closely related to the construct of the PROM)?

Elaboration: *Q3.1 is an alternative assessment for Q3. When the authors did not report the correlation between anchor and PROM. We assess Q3.1 instead. Q3.1is rated on a 5-point adjectival scale with response options for 'definitely closely related'; 'to a great extent'; 'not so much'; 'definitely not related' and 'impossible to tell' (lowest credibility). Raters base their judgements on the extent of the anticipated associations between the PROM construct and the anchor construct: the closer the anticipated association, the higher the rating. For example, when both the anchor and the PROM measure pain, we will rate Q3.1 as 'definitely closely related'. Our parallel work provided detailed principles addressing the assessments of Q3.1 for the most frequently used anchors: transition ratings, measures of satisfaction, other PROMs, and clinical measures [4].*

Q4. Is the MID precise?

Elaboration: *Precision around the MID estimate is quantified by the difference between the point estimate and the boundaries (lower and upper) of the CI and expressed as a percentage. In many*

cases, the authors may not report any measure of variability (SD, SE, CI, range, etc.). In these situations, we ask that you consider the sample size used to estimate the MID. Q4 is rated on a 5-point adjectival scale with response options for 'definitely yes' (<10% or ≥ 200 patients; highest credibility); 'to a great extent' (11-25% or 150-199 patients); 'not so much' (26-49% or 100-149 patients); 'definitely no'(≥ 50% or <100 patients) and 'impossible to tell' (lowest credibility).

Q5. Does the threshold or difference between groups on the anchor used to estimate the MID reflect a small but important difference?

Elaboration: *Q5 is rated on a 5-point adjectival scale with response options for 'definitely yes' (highest credibility); 'to a great extent'; 'not so much'; 'definitely no' and 'impossible to tell' (lowest credibility). We judge whether the chosen threshold or groups compared on the anchor reflect a small (rather than moderate or large) but important difference. In addition, whether the chosen method of analysis calculates an MID need to be determined. A framework for make these judgements include: '1. What is the original scale of the anchor, and it is transformed on 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? Our previous report has more detailed description for the judgement'*[3].

2. Extension criteria exclusively for transition rating anchors

Q6. Is the amount of elapsed time between baseline and follow-up measurement for MID estimation optimal?

Elaboration: *Q6 is rated on a 5-point adjectival scale with response options for 'definitely yes' (≤ 4 weeks or 1 month; highest credibility); 'to a great extent' (1 to 2 months); 'not so much' (>2 to 3 months); 'definitely no' (>3 months) and 'impossible to tell' (lowest credibility). This criterion is to capture the recall bias of transition anchors. As time extends into months, patients are more likely to confuse change over time with current status.*

Judgments for Q7-9 requires knowledge of the directional characteristics of the patient reported outcome measure and transition scale. In the following elaboration, we assume that higher values on the anchor and PROM represent the same state (i.e., both represent a better or worse condition).

Q7. Does the transition item have a satisfactory correlation with the PROM score at follow-up?

Elaboration: *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 [5]. 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. Q7 is rated on a 5-point adjectival scale with response options for 'definitely yes' (≥0.2, highest credibility); 'to a great extent' (0.1 to 0.2); 'not so much' (0 to 0.1); 'definitely no' (negative correlation) and 'impossible to tell' (lowest credibility).*

Q8. Does the transition item correlate with the PROM score at baseline?

Elaboration: *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 [5]. 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. Q8 is rated on a 5-point adjectival scale with response options for 'definitely yes' (negative correlation, highest credibility); 'to a great extent' (0 to 0.1); 'not so much' (0.1 to 0.2); 'definitely no' (≥0.2) and 'impossible to tell' (lowest credibility).*

Q9. 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?

Elaboration: *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 transition item with the PROM change score that is appreciably greater than the correlation of the transition item with the PROM score at follow-up is less likely to reflect current status, and thus more confidence in the MID estimate[3]. Q9 is rated on a 5-point adjectival scale with response options for 'definitely*

yes' (≥0.2, highest credibility); 'to a great extent' (0.1 to 0.2); 'not so much' (0 to 0.1); 'definitely no' (negative correlation) and 'impossible to tell' (lowest credibility).

B. Select the most credible MIDs (*Figure 5, box 1-4*)

The selection process begins with identifying the most credible MIDs (*Figure 5, box 1-4*). To do so, we apply the credibility instrument described above and count the number of the five core criteria met and select the MID estimates that satisfy the greatest number of criteria. The most credible MIDs are those with the highest ratings across the 5 core credibility criteria (i.e., the highest credibility rank, *Appendix 8*).

The best MIDs are those meeting five 'definitely yes/yes/definitely closely related' for the five core criteria (*Figure 5, box 1*). To arrive at the most credible MID estimates, however, we may progressively relax the criteria until they find available MIDs with the highest credibility rank (*Appendix 8*). The detailed process is as follows (*Figure 5, box 1-4*): first, choose MID estimates with 5 '*definitely yes/yes/definitely closely related'* across the credibility core criteria (Q1, Q2, Q4, Q5 and either correlation criterion (Q3) or construct criterion (Q3.1) if correlation coefficient is not reported) (i.e., *Figure 5, box 1*); if, however, there are no estimates meeting 5 '*definitely yes/yes/definitely closely related*' ratings, we relax our definition of 'most credible' MIDs, and include MIDs that are rated as '*definitely yes/yes/definitely closely related*' or *'to a great extent*' across the 5 core criteria. If there are no MIDs meeting this relaxed definition, we further relax our criteria to select the estimates in lower credibility rank (see *Appendix 8* for the credibility ranks, higher rank means higher credibility), that is, the estimates with ratings of '*definitely yes/yes/definitely closely related*' in 4 core criteria. If not available again, we select MIDs with ratings of '*definitely yes/yes/definitely closely related*'' or '*to a great extent'* in 4 core criteria and so on (*Appendix 8*). Of note, when MIDs with higher rank are available, we would not go down the ranking system to select the MIDs with lower rank. For example, if we have one estimate at rank 1, this estimate would be the only one we will use.

Then, we will check the **consistency*** of the most credible MID estimates (*Figure 5, box 5*). MIDs are considered consistent with one another if 80% of the estimates lie within an absolute value of

10% of the PROM score above or below their median (i.e., within a range of 20% of the median). Then, we will compare the distribution of the most credible MIDs to the distribution of all available MIDs (i.e., all MIDs irrespective of their credibility) *(Figure 5, box 9*). The consistency and distribution of the MIDs determines the next steps of the selection process (*Figure 5, box 6,7,12,13*).

C. When the most credible estimates are consistent with each other (Figure 5, box 7)

When the most credible MIDs are consistent with each other (*Figure 5, box 7),* we will check if the median of the most credible MIDs is near the median of all MIDs reported for a given PROM (*Figure 5, box 9*).

If the absolute difference between the median of the most credible MID and the median of the whole distribution is less than an absolute value of 10% of the PROM score, the median of the most credible MID is considered as **'near*'** the median of all MIDs reported for the PROM.

When the median of the most credible MIDs is near the median of all MIDs reported for a given PROM (*Figure 5, box 13*), we can be confident that **the median of the most credible MID represents the optimal MID** (*Figure 5, box 19*). In this scenario, the optimal MID is applicable to all contexts. Our worked examples, VAS-pain and SF-36-MCS, apply to this scenario (*Appendix 9*).

D. When the most credible estimates are not consistent with each other (Figure 5, box 6)

When the most credible MIDs are not consistent with each other, which indicates excessive variability may exist, we will try to explain the variability by further consideration of the credibility criteria.

D.1 Correlation criterion (Q3)

The most important credibility criterion addresses anchor validity: the correlation between the anchor and the PROM of interest [3]. We, however, at the very beginning, treat all the 5 core criteria

equally, including the correlation criterion (Q3) and its alternative construct proximity criterion (Q3.1). When the most credible MIDs are not consistent, we will therefore prioritize the correlation criterion (Q3) to explain the remaining variability and among the most credible MIDs, select the MIDs with high correlations with the anchor (r≥0.5) (*Figure 5, box 8*).

If such estimates are available (*Figure 5, box 11*), we will check the consistency of these MIDs (*Figure 5, box 15,17,18)* and compare the distribution of these MIDs with the distribution of all available MIDs for the PROM of interest (*Figure 5, box 22,24,25).*

When the MID estimates with high correlation with the anchors are consistent (*Figure 5, box 18*) and their median is near the median of all MIDs (*Figure 5, box 25*), **the median of the MID estimates with high correlation with the anchors represents the optimal MID** (*Figure 5, box 28*).

D.2 Recall period criterion (Q6)

If, however, substantial variability among the MID estimates with high correlations with the anchor (r≥0.5) remains (*Figure 5, box 17*), or if no MIDs with high correlations are available *(Figure 5, box 10)*, because MIDs were often estimated on transition rating anchors (more than half in the MID inventory), investigators could further consider the additional extension credibility criteria that are exclusively important to address the validity of transition rating anchors [3 5]. We suggest assessing the recall period (Q6, see above) for the transition rating anchors because except follow-up length, the relevant data for the other three additional criteria were rarely reported [3]. We will remove the estimates anchored to transition ratings with a long recall period (>4 weeks) (*Figure 5, box 14)*.

We then determine if the remaining most credible estimates are consistent (*Figure 5, box 16,20,21)* and compare the distribution of the remaining most credible MIDs with the distribution of all available MIDs for the PROM (*Figure 5, box 23,26,27).*

When the remaining most credible MID estimates are consistent (*Figure 5, box 21*) and their median is near the median of all MIDs (*Figure 5, box 27*), **the median of the remaining most credible MID estimates represents the optimal MID** (*Figure 5, box 29*).

If, however, among the most credible estimates, no MIDs are anchored to transition ratings, or all the most credible estimates anchored to transition ratings are with long recall period (>4 weeks), we will skip *Step 1.D2 (i.e., the dotted boxes in Figure 5, box 14,16,20,21,23,26,27,29)*.

D.3 Notes for further consideration of the credibility criteria

The further consideration of the credibility criteria regarding the anchor validity (D.1 and D.2), however, may not be necessary. Because there may be few estimates with a high correlation of 0.5 between the PROM and the anchor; it is possible that all the estimates using transition anchors have long recall period; more often, the most credible estimates would appear to be consistent. For example, in our worked examples (*Appendix 9, 10*), we did not further consider the correlation and recall period criteria. After applying the five core credibility assessments, the most credible MIDs were consistent (*Appendix 9, 10*).

**NOTE*
1. '*Consistency*': Consistency is defined as 80% of the MID estimates lying within an absolute value of 10% of the PROM score above or below the median (i.e., within a range of 20% of the median).
For example: WOMAC-function (0-100): the most credible MIDs are ***26.54,33.5 and 23.***
The median of the most credible MIDs is 26.54
Then the consistency range would be 16.54-36.54
In the example, 80% MIDs fall in the consistency range and we consider the most credible estimates are consistent.

2. '*Near':* The absolute difference between the median of the selected MID estimates and the median of the whole distribution is less than an absolute value of 10% of the PROM score.
For example: WOMAC-function (0-100)
median of the most credible MIDs=26.54
median of the overall MIDs=12
26.54-12=14.54 >10
In the example, the median of the most credible is not near the median of the overall.

Step 2: Explore contextualized factors.

After *Step 1*, we may end up with the following situations, which support the MIDs for a PROM could be context-dependent. In this case, when selecting and applying optimal MID, we will require further exploration seeking contextualization as an explanation of MID variability (*Step 2*). The potential contextualized factors that deserve a consideration could come from the suggestions of previous researchers, including intervention (e.g., surgical vs conservative treatments), patient condition (e.g., knee vs hip osteoarthritis), baseline disease severity, patient age (e.g., young vs old), follow-up length, socioeconomic status, geography, and sex [2].

Situation 1: The most credible MIDs (see *Step 1. B; Figure 5, box 12*) or the newly identified most credible MIDs (see *Step 1.C, D*; *Figure 5, box 24, 26*) are consistent but their median differs considerably from the median of the whole distribution (see WOMAC-pain (*Appendix 10*) and KOOS-Qol examples (*Appendix 9)*).

Situation 2: The most credible MIDs (see *Step 1*.B*; Figure 5, box 6*) or the newly identified MIDs are inconsistent (see *Step 1.C, D*; *Figure 5, box 17, 20*).

In *Step 2*, we place an order of using the available estimates for the exploration:

1. when there are a **sufficient number**[#] of the most credible MIDs, use the most credible estimates (i.e., the estimates from *Figure 5, box 5*) to explore the MID variability (*Figure 5, box 31).*

2. when there are insufficient number of the most credible MID estimates, use all available MID estimates to explore the MID variability (*Figure 5, box 35).*

Based on the contextualized factors (usually binary factor), we categorize the MID estimates into two groups (e.g., surgical group vs non-surgical group), and to test whether the MID variability

could be explained by the factors, we compare the medians of the two groups using Wilcoxon Rank Sum Test with a threshold *p* value of *0.10* (*Figure 5, box 34, 41*).

The final choice and application of the optimal MID will be determined by whether the contextualized factors can explain the MID variability.

a. If the search for contextualized factors yields an explanation for the variability in MIDs (*Figure 5, box 36, 42*), **the optimal MID will be context-dependent**, **and we will identify one optimal MID for each context** (*Figure 5, box 40, 44*). The application of the optimal MID will also be context- dependent.

For instance, the MID estimate of a particular PROM is different for surgical and non-surgical interventions. In this case, there are two optimal MIDs for the PROM: one MID that should be used exclusively in the context of surgical interventions, and another to be used in the context of non-surgical intervention. (see WOMAC-pain (*Appendix 10*) and KOOS-Qol examples (*Appendix 9*).

b. If no contextual explanation is found (*Figure 5, box 37, 39, 43*), we will still select the median of the most credible estimates (*Figure 5, box 5*) among all available MIDs as the optimal MID. In this case, the optimal MID is applicable to all contexts (*Figure 5, box 45*).

[#]*NOTE*
Analogously to subgroup analysis, we will use the same threshold for '*sufficient number*' of estimates. That is, at least 3 estimates per contextualized factor.
For example, to explore the impact of surgical versus non-surgical interventions on MID variability, there should be at least 6 estimates.

Reference

1. Zhang Y, Zhang S, Thabane L, et al. Although not consistently superior, the absolute approach to framing the minimally important difference has advantages over the relative approach. *J Clin Epidemiol* 2015;68(8):888-94. doi: 10.1016/j.jclinepi.2015.02.017
2. Wang Y, Devji T, Qasim A, et al. A systematic survey identified methodological issues in studies estimating anchor-based minimal important differences in patient-reported outcomes. *J Clin Epidemiol* 2022;142:144-51. doi: 10.1016/j.jclinepi.2021.10.028
3. 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. doi: 10.1136/bmj.m1714
4. Wang Y, Devji T, Carrasco-Labra A, et al. An extension MID credibility item addressing construct proximity is a reliable alternative to correlation item. *J Clin Epidemiol* 2023 doi: 10.1016/j.jclinepi.2023.03.001
5. Guyatt GH, Norman GR, Juniper EF, et al. A critical look at transition ratings. *J Clin Epidemiol* 2002;55(9):900-8. doi: 10.1016/s0895-4356(02)00435-3

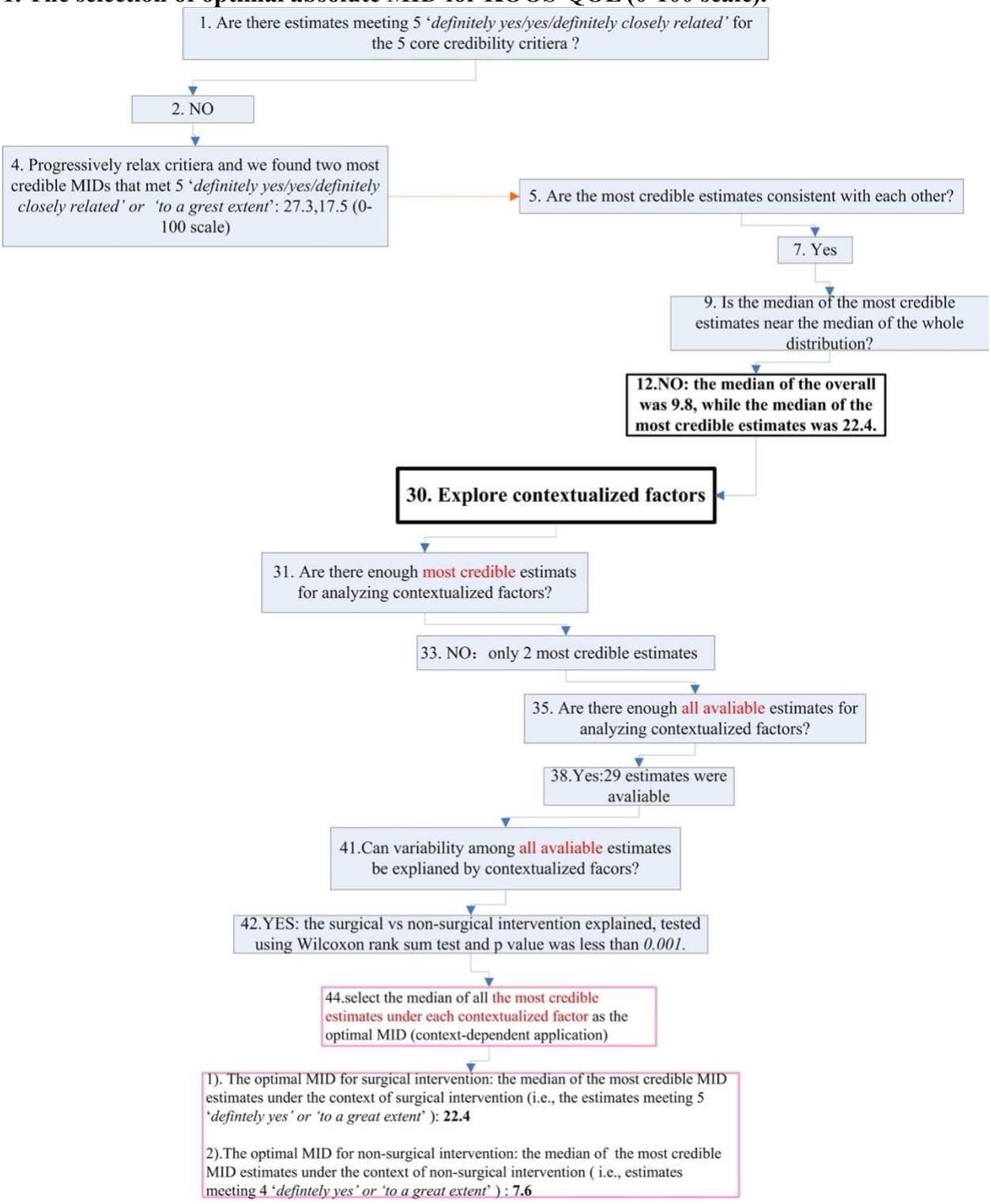
Appendix 8. The ranks for MID credibility.

Rank 1.	5 ‘definitely yes/yes/definitely closely related’
Rank 2.	5 ‘definitely yes/yes/definitely closely related’ or ‘to a great extent’ (any combinations).
Rank 3.	4 ‘definitely yes/yes/definitely closely related’
Rank 4.	4 ‘definitely yes/yes/definitely closely related’ or ‘to a great extent’ (any combinations).
Rank 5.	3 ‘definitely yes/yes/definitely closely related’
Rank 6.	3 ‘definitely yes/yes/definitely closely related’ or ‘to a great extent’ (any combinations).
Rank 7.	2 ‘definitely yes/yes/definitely closely related’
Rank 8.	2 ‘definitely yes/yes/definitely closely related’ or ‘to a great extent’ (any combinations).
Rank 9.	1 ‘definitely yes/yes/definitely closely related’
Rank 10.	1 ‘definitely yes/yes/definitely closely related’ or ‘to a great extent’ (any combinations).
Rank 11.	0 ‘definitely yes/yes/definitely closely related’ or ‘to a great extent’ (any combinations).

*The rank is based on the 5 core criteria in the MID credibility assessment instrument. Higher rank means higher credibility.

Appendix 9. The selection process for Knee injury and Osteoarthritis Outcome Score-quality of life (KOOS-Qol) and Pain visual analogue scale (VAS-pain) and the 36-item Short Form Survey-mental component summary (SF-36-MCS) (up to 2018).

1. The selection of optimal absolute MID for KOOS-QOL (0-100 scale).



Elaboration:

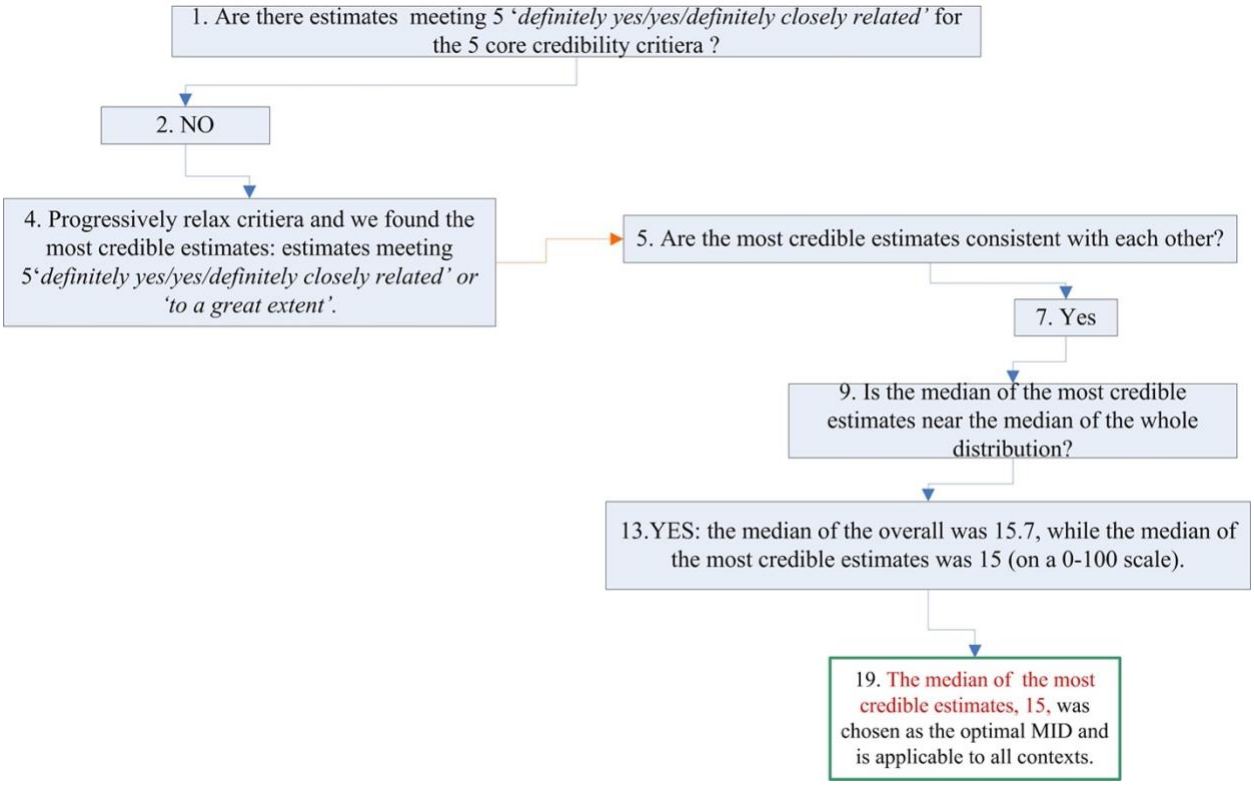
We selected the optimal MID from the absolute MID estimates. There were 29 absolute MID estimates (up to 2018) for KOOS-QOL (0-100), ranging from *3* to *27.3*. The median of all the MIDs was *9.8*. After assessing the five core credibility criteria, no estimates met 5'*definitely yes/yes/definitely closely related*'. We therefore relaxed our criteria and found that the MIDs with highest rank available, referring to the most credible MIDs, were those meeting 5 '*definitely yes/yes/definitely closely related'* or '*to a great extent'*. We then stopped searching for MIDs in lower ranks. The most credible MIDs were *27.3* [1] *and 17.5* [1], which were consistent. We thus went to check if the median of these most credible MIDs (i.e., *22.4*), is near the median of all the available MIDs (i.e., *9.8*). Because the KOOS-QOL was on a 0-100 scale, the absolute difference between the two medians were above a score of *10*. We considered the two medians were not near, which indicated that it was possible that contextualized factors would further impacting the MID variability.

We then went to explore the contextualized factors impacting the MID variability using all available MIDs because we only had 2 most credible MIDs. We found that the intervention (surgical vs non-surgical interventions) has impacted the MID variability for KOOS-QOL ($p<0.001$). Therefore, the selection and application of the optimal MID for KOOS-QOL was context-dependent—the optimal MID differed depending on whether the patients were undergoing surgical intervention.

Under the context of surgical intervention, the point estimates for the most credible MIDs that met 5 '*definitely yes/yes/definitely closely related'* or '*to a great extent'* were *27.3* [1] and *17.5* [1]. We took the median—*22.4*—as the optimal MID, which was applicable to the context of surgical intervention.

Under the context of non-surgical intervention, we did not find MIDs met neither 5 '*definitely yes/yes/definitely closely related'* nor 5 '*definitely yes/yes/definitely closely related'* or '*to a great extent'*. But the MIDs met 4 '*definitely yes/yes/definitely closely related*' or '*to a great extent'* were available, which were considered as the most credible estimates. There were 11 most credible estimates that met 4 '*definitely yes/yes/definitely closely related*' or '*to a great exte*nt', including *9.9* [2]*,6.6* [2]*,9.4* [3]*, 3* [4] *,3* [4] *,11.62* [4]*, 7.6* [4]*, 3* [4] *,3* [4]*, 9.8* [4]*, 7.8* [4]. We took the median—*7.6*—as the optimal MID, which was applicable to the context of non-surgical intervention.

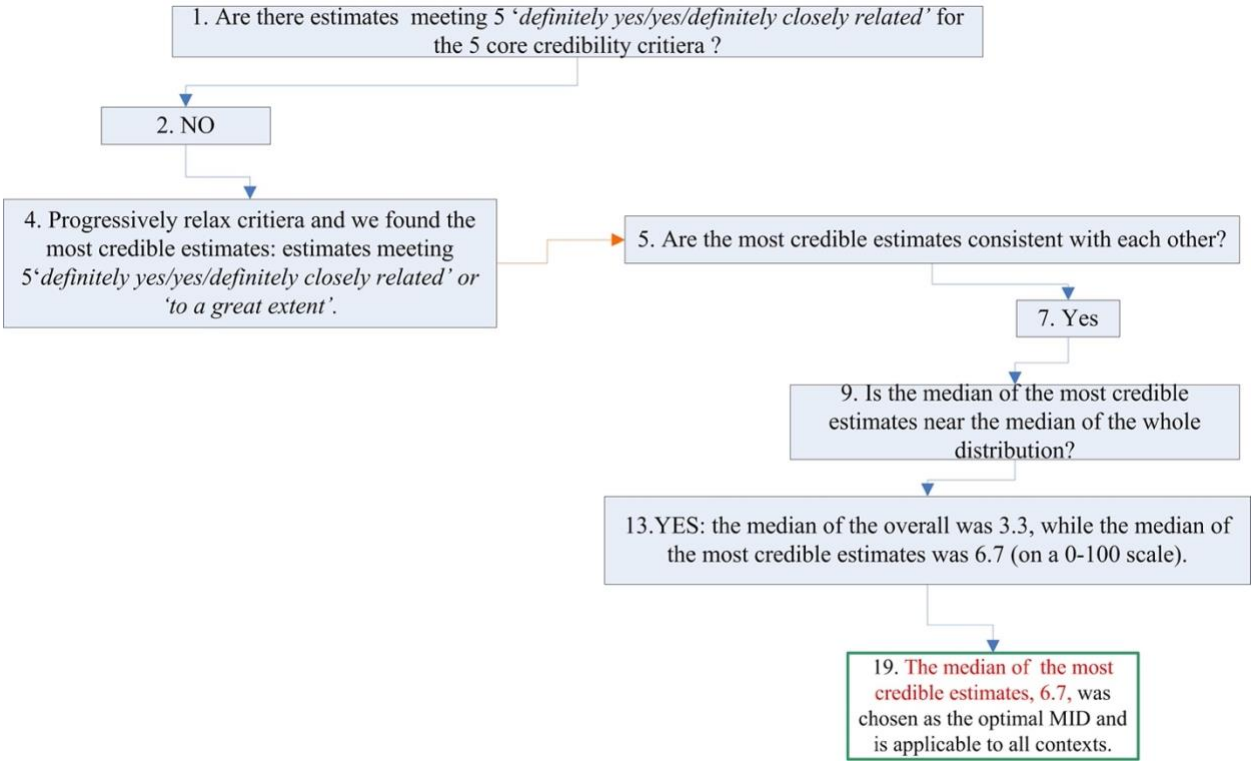
2. The selection of optimal absolute MID for VAS-pain (0-100 scale).



Elaboration:

We selected the optimal MID from the absolute MID estimates. There were 118 absolute MID estimates (up to 2018) for VAS-pain (0-100), ranging from *0.95* to *88.7*. The median of all the MIDs was *15.7*. After assessing the five core credibility criteria, no estimates met 5'*definitely yes/yes/definitely closely related*'. We therefore relaxed our criteria and found that the MIDs with highest rank available, referring to the most credible MIDs, were those meeting 5 '*definitely yes/yes/definitely closely related'* or '*to a great extent'*. We then stopped searching for MIDs in lower ranks. The most credible MIDs were *17* [5]*, 13.5* [6]*, 12* [7]*, 15* [8]*, 13* [8]*,16* [9]*, 14* [10]*,17* [11]*, 20.4* [12], which, by definition, were consistent with each other. We thus went to check if the median of these most credible MIDs (i.e., *15*), is near the median of all the available MIDs (i.e., *15.7*). Because the VAS-pain was on a 0-100 scale, the absolute difference between the two medians were within a score of *10*. We thus considered the two medians were near, which indicated that there were no contextualized factors further impacting the MID variability. Therefore, the MIDs of VAS-pain was considered as non-contextualized. We took the median of the most credible MIDs—*15*—as the optimal MID, which was applicable to all contexts.

3. The selection of optimal absolute MID for SF-36-MCS (0-100 scale).



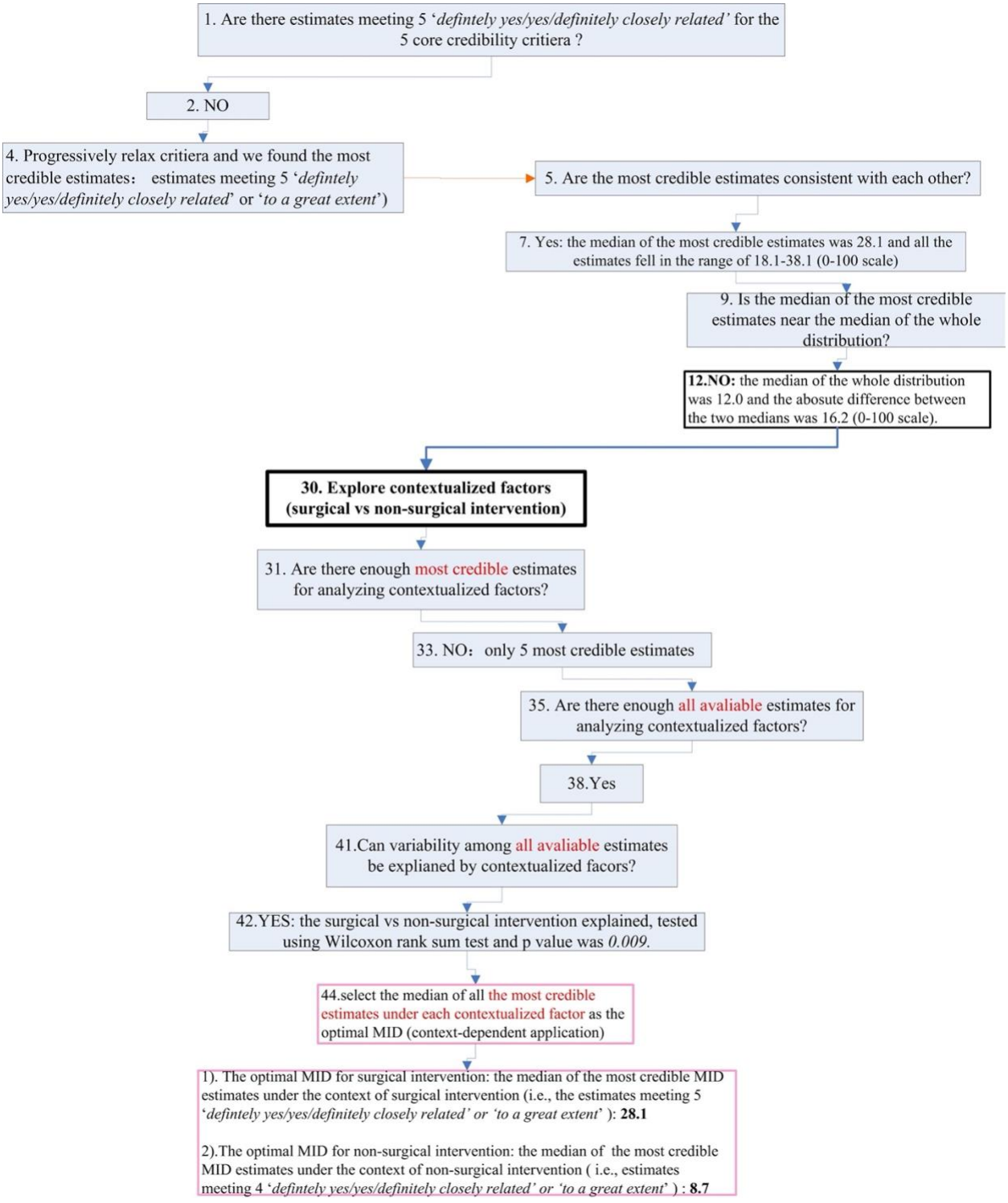
Elaboration:

We selected the optimal MID from the absolute MID estimates. There were 45 absolute MID estimates (up to 2018) for SF-36-MCS ranging from *0.11* to *49.5*. The median of all the MIDs was *3.3* (on a 0-100 scale). After assessing the five core credibility criteria, no estimates met 5'*definitely yes/yes/definitely closely related*'. We therefore relaxed our criteria and found the MIDs with highest rank available, referring to the most credible MIDs, were those meeting 5 '*definitely yes/yes/definitely closely related'* or '*to a great extent'*. We then stopped searching for MIDs in lower ranks. The most credible MIDs were*, 7.25* [13]*, 6.7* [13]*, 5.0* [14], which were consistent. We thus went to check if the median of these most credible MIDs (i.e., *6.7*), is near the median of all the available MIDs (i.e., *3.3*). Because the SF-36-MCS was on a 0-100 scale, the absolute difference between the two medians were within a score of *10*. We considered the two medians were near, which indicated that there were no contextualized factors further impacting the MID variability. Therefore, the MIDs of SF-36-MCS was considered as non-contextualized. We took the median of the most credible MIDs—*6.7*—as the optimal MID, which was applicable to all contexts.

References

1. Ingelsrud LH, Terwee CB, Terluin B, et al. Meaningful Change Scores in the Knee Injury and Osteoarthritis Outcome Score in Patients Undergoing Anterior Cruciate Ligament Reconstruction. *Am J Sports Med* 2018;46(5):1120-28. doi: 10.1177/0363546518759543
2. Crossley KM, Macri EM, Cowan SM, et al. The patellofemoral pain and osteoarthritis subscale of the KOOS (KOOS-PF): development and validation using the COSMIN checklist. *Br J Sports Med* 2018;52(17):1130-36. doi: 10.1136/bjsports-2016-096776
3. Huang CC, Chen WS, Tsai MW, et al. Comparing the Chinese versions of two knee-specific questionnaires (IKDC and KOOS): reliability, validity, and responsiveness. *Health Qual Life Outcomes* 2017;15(1):238. doi: 10.1186/s12955-017-0814-6
4. Mills KA, Naylor JM, Eyles JP, et al. 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(2):395-404. doi: 10.3899/jrheum.150398
5. Bird SB, Dickson EW. Clinically significant changes in pain along the visual analog scale. *Ann Emerg Med* 2001;38(6):639-43. doi: 10.1067/mem.2001.118012
6. Lopez BL, Flenders P, Davis-Moon L, et al. Clinically significant differences in the visual analog pain scale in acute vasoocclusive sickle cell crisis. *Hemoglobin* 2007;31(4):427-32. doi: 10.1080/03630260701587810
7. Kelly AM. The minimum clinically significant difference in visual analogue scale pain score does not differ with severity of pain. *Emerg Med J* 2001;18(3):205-7. doi: 10.1136/emj.18.3.205
8. Gallagher EJ, Liebman M, Bijur PE. Prospective validation of clinically important changes in pain severity measured on a visual analog scale. *Ann Emerg Med* 2001;38(6):633-8. doi: 10.1067/mem.2001.118863
9. Gallagher EJ, Bijur PE, Latimer C, et al. Reliability and validity of a visual analog scale for acute abdominal pain in the ED. *Am J Emerg Med* 2002;20(4):287-90. doi: 10.1053/ajem.2002.33778
10. Bijur PE, Latimer CT, Gallagher EJ. Validation of a verbally administered numerical rating scale of acute pain for use in the emergency department. *Acad Emerg Med* 2003;10(4):390-2. doi: 10.1111/j.1553-2712.2003.tb01355.x
11. Mark M, Au T, Choi Y, et al. The minimum clinically significant difference in visual analogue scale pain score in a local emergency setting. *Hong Kong Journal of Emergency Medicine* 2009;16(4):233-36.
12. Ward MM, Guthrie LC, Alba MI. Clinically important changes in individual and composite measures of rheumatoid arthritis activity: thresholds applicable in clinical trials. *Ann Rheum Dis* 2015;74(9):1691-6. doi: 10.1136/annrheumdis-2013-205079
13. Carlson ML, Tveiten Ø V, Yost KJ, et al. The Minimal Clinically Important Difference in Vestibular Schwannoma Quality-of-Life Assessment: An Important Step beyond P < .05. *Otolaryngol Head Neck Surg* 2015;153(2):202-8. doi: 10.1177/0194599815585508
14. Skolasky RL, Albert TJ, Maggard AM, et al. Minimum clinically important differences in the Cervical Spine Outcomes Questionnaire: results from a national multicenter study of patients treated with anterior cervical decompression and arthrodesis. *J Bone Joint Surg Am* 2011;93(14):1294-300. doi: 10.2106/jbjs.J.01136

Appendix 10. The selection process of optimal MID for WOMAC-pain.



Appendix 11. Characteristics of all anchor-based MIDs for the WOMAC-pain (up to 2018).

Characteristic	N (total=67)
Anchor-type	
Transition rating	64
Satisfaction rating	4
PROM score range	
0-10	3
0-20	1
0-50	6
0-100	56
Not reported	1
MID direction	
improvement	63
deterioration	4
MID expressed in absolute terms	45
MID expressed relative to baseline scores	22
MID estimation contexts	
intervention	21
surgical intervention	46
non-surgical intervention	
Patient condition	
Knee related complaints	36
Hip related complaints	16
Knee or hip related complaints	15
MID estimation follow-up length	
≤6 months	51
>6 months	16
MID credibility rated as '*definitely yes*'*	
Q1	67
Q2	56
Q3	0
Q3.1	8
Q4	5
Q5	8

NOTE:
*For credibility criterion Q1, the highest rating is '*yes*' and for credibility criterion *Q3.1*, the highest rating is '*definitely closely related'*.

Appendix 12. All relevant data of WOMAC-pain absolute MIDs for the selection.

Ref NO.[&]	Contexts of the study		PROM scale		MID information			Credibility assessment[#]					
	Condition	Intervention	Lower Limit	Upper Limit	MID point estimate	Absolute value of the point estimate[*]	MID follow-up length (months)	Core Cred Q1	Core Cred Q2	Core Cred Q3/Q3.1	Core Cred Q4	Core Cred Q5	Credibility Ranking
1	Femoroacetabular impingement	arthroscopic surgery with labral preservation or limited anterolateral open surgery with labral resection	0	100	28	28	6	1	3	2	1	2	4
2	hip osteoarthritis, knee osteoarthritis	primary or revision total hip or knee replacement	0	20	3.3	16.5	6	1	3	3	0	0	5
3	Hip complaints	usual care	0	100	0	0	3	1	3	1	0	2	6
3	Hip complaints	usual care	0	100	7.4	7.4	3	1	3	1	0	2	6
3	hip osteoarthritis	behavioral-graded activity or usual care	0	100	11.2	11.2	3	1	3	1	1	2	6
3	hip osteoarthritis	behavioral-graded activity or usual care	0	100	12.1	12.1	3	1	3	1	0	2	6
3	hip osteoarthritis	Total hip replacement	0	100	8.3	8.3	6	1	3	1	0	3	5
3	hip osteoarthritis	Total hip replacement	0	100	22.4	22.4	6	1	3	1	0	3	5
3	Knee pain	usual care	0	100	7.4	7.4	18	1	3	0	0	1	7
3	Knee pain	usual care	0	100	3.5	3.5	18	1	3	0	0	1	7
3	knee complaints	usual care	0	100	11.8	11.8	12	1	3	2	1	2	4
3	knee complaints	usual care	0	100	12.9	12.9	12	1	3	2	0	2	4
3	knee complaints	usual care	0	100	5.3	5.3	3	1	3	1	0	2	6
3	knee complaints	usual care	0	100	11	11	3	1	3	1	0	2	6
3	knee osteoarthritis	behavioral-graded activity or usual care	0	100	10.4	10.4	3	1	3	1	1	2	6
3	knee osteoarthritis	behavioral-graded activity or usual care	0	100	0.1	0.1	3	1	3	1	0	2	6
3	knee osteoarthritis	total knee replacement	0	100	13.3	13.3	6	1	3	1	0	3	5
3	knee osteoarthritis	total knee replacement	0	100	29.4	29.4	6	1	3	1	1	3	5
4	Osteoarthritis of the Lower Extremities	inpatient rehabilitation intervention, which consisted of: passive physical therapy, such as electrotherapies,	0	100	6.4	6.4	3	1	3	2	0	2	4

		hydrotherapies, thermotherapies, massage, and others, and of especially active physical therapy to strengthen and stretch the musculature and passive structures, and to reestablish regular joint mobility.											
4	Osteoarthritis of the Lower Extremities	inpatient rehabilitation intervention, which consisted of: passive physical therapy, such as electrotherapies, hydrotherapies, thermotherapies, massage, and others, and of especially active physical therapy to strengthen and stretch the musculature and passive structures, and to reestablish regular joint mobility.	0	10	-0.83	8.3	3	1	3	2	1	2	4
5	hip osteoarthritis	total hip replacement	NR	NR	14.66	14.66	6	1	2	2	4	4	6
6	Hip osteoarthritis	Total hip replacement	0	100	29.26	29.26	6	1	3	2	2	2	2
7	articular cartilage defect of the knee	Surgical treatment (shaving, drilling, autologous chondrocyte implantation (ACI), abrasion arthroplasty, microfracture, and cell therapy (or))	0	100	17.5	17.5	6	1	3	2	0	1	6
8	Knee osteoarthritis	total knee replacement	0	100	29.9	29.9	12	1	3	2	3	2	2
8	Knee osteoarthritis	total knee replacement	0	100	20.5	20.5	12	1	3	2	3	2	2
8	Knee osteoarthritis	total knee replacement	0	100	28	28	12	1	3	1	2	0	6
8	Knee osteoarthritis	total knee replacement	0	100	25.2	25.2	12	1	3	1	2	2	4
8	Knee osteoarthritis	total knee replacement	0	100	28.1	28.1	12	1	3	2	3	2	2
8	Knee osteoarthritis	total knee replacement	0	100	23.5	23.5	12	1	3	2	3	2	2
8	Knee osteoarthritis	total knee replacement	0	100	25.6	25.6	12	1	3	1	2	0	6

8	Knee osteoarthritis	total knee replacement	0	100	27.5	27.5	12	1	3	1	2	2	4
9	hip osteoarthritis, knee osteoarthritis	physical therapy or usual medical care	0	50	2.5	5	2.25	1	1	2	0	2	6
9	hip osteoarthritis, knee osteoarthritis	physical therapy or usual medical care	0	50	4.5	9	2.25	1	1	2	0	2	6
9	hip osteoarthritis, knee osteoarthritis	physical therapy or usual medical care	0	50	5.5	11	2.25	1	1	2	0	2	6
9	hip osteoarthritis, knee osteoarthritis	physical therapy or usual medical care	0	50	5	10	2.25	1	1	2	0	2	6
9	hip osteoarthritis, knee osteoarthritis	physical therapy or usual medical care	0	50	9	18	2.25	1	1	2	0	2	6
9	hip osteoarthritis, knee osteoarthritis	physical therapy or usual medical care	0	50	11	22	2.25	1	1	2	0	2	6
10	knee osteoarthritis	Rehabilitation intervention	0	100	13.51	13.51	3	1	3	3	1	2	4
10	knee osteoarthritis	Rehabilitation intervention	0	100	8.74	8.74	3	1	3	3	0	2	4
10	knee osteoarthritis	Rehabilitation intervention	0	100	15	15	3	1	3	3	1	2	4
10	knee osteoarthritis	Rehabilitation intervention	0	100	8.74	8.74	3	1	3	3	0	2	4
10	knee osteoarthritis	Rehabilitation intervention	0	100	7.09	7.09	3	1	3	3	0	2	4
11	hip osteoarthritis, knee osteoarthritis	non-steroidal anti-inflammatory drug (NSAID)	0	100	9	9	1	1	1	2	1	2	6
12	hip osteoarthritis, knee osteoarthritis	multimodal conservative treatment (comprised education, physical therapy, step-up analgesics, and advice on weight reduction if needed)	0	100	-4.1	4.1	3	1	2	2	4	2	4
13	hip fractures	rehabilitation program	0	100	35	35	2	1	3	1	1	1	7

Notes:

[&]References:

1. Impellizzeri FM, Mannion AF, Naal FD, et al. The early outcome of surgical treatment for femoroacetabular impingement: success depends on how you measure it. *Osteoarthritis Cartilage* 2012;20(7):638-45. doi: 10.1016/j.joca.2012.03.019

2. Terwee CB, Roorda LD, Knol DL, et al. Linking measurement error to minimal important change of patient-reported outcomes. *J Clin Epidemiol* 2009;62(10):1062-7. doi: 10.1016/j.jclinepi.2008.10.011

3. Terwee CB, Roorda LD, Dekker J, et al. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol* 2010;63(5):524-34. doi: 10.1016/j.jclinepi.2009.08.010
4. Angst F, Aeschlimann A, Michel BA, et al. Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities. *J Rheumatol* 2002;29(1):131-8.
5. Quintana JM, Aguirre U, Barrio I, et al. Outcomes after total hip replacement based on patients' baseline status: what results can be expected? *Arthritis Care Res (Hoboken)* 2012;64(4):563-72. doi: 10.1002/acr.21570
6. Quintana JM, Escobar A, Bilbao A, et al. Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement. *Osteoarthritis Cartilage* 2005;13(12):1076-83. doi: 10.1016/j.joca.2005.06.012
7. Greco NJ, Anderson AF, Mann BJ, et al. Responsiveness of the International Knee Documentation Committee Subjective Knee Form in comparison to the Western Ontario and McMaster Universities Osteoarthritis Index, modified Cincinnati Knee Rating System, and Short Form 36 in patients with focal articular cartilage defects. *Am J Sports Med* 2010;38(5):891-902. doi: 10.1177/0363546509354163
8. Escobar A, García Pérez L, Herrera-Espiñeira C, et al. Total knee replacement; minimal clinically important differences and responders. *Osteoarthritis Cartilage* 2013;21(12):2006-12. doi: 10.1016/j.joca.2013.09.009
9. Abbott JH, Hobbs C, Gwynne-Jones D. The ShortMAC: Minimum Important Change of a Reduced Version of the Western Ontario and McMaster Universities Osteoarthritis Index. *J Orthop Sports Phys Ther* 2018;48(2):81-86. doi: 10.2519/jospt.2018.7676
10. Angst F, Aeschlimann A, Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. *J Clin Epidemiol* 2017;82:128-36. doi: 10.1016/j.jclinepi.2016.11.016
11. Bellamy N, Hochberg M, Tubach F, et al. Development of multinational definitions of minimal clinically important improvement and patient acceptable symptomatic state in osteoarthritis. *Arthritis Care Res (Hoboken)* 2015;67(7):972-80. doi: 10.1002/acr.22538
12. Mahler E, den Broeder AA, Woodworth TG, et al. How should worsening in osteoarthritis be defined? Development and initial validation of preliminary criteria for clinical worsening in knee and hip osteoarthritis. *Scand J Rheumatol* 2017;46(5):396-406. doi: 10.1080/03009742.2016.1235226
13. Monticone M, Ambrosini E, Secci C, et al. Responsiveness and Minimal Important Changes of the Western Ontario and McMaster Universities Osteoarthritis Index in Subjects Undergoing Rehabilitation Following Hip Fracture. *Am J Phys Med Rehabil* 2017;96(5):321-26. doi: 10.1097/phm.0000000000000609

*Taking the absolute value of the point estimates and all transformed into a 0-100 scale.

#See Appendix 7 for all credibility core criteria. Highlighted yellow columns refers to that the correlation coefficients were not reported, and we assessed Q3.1 instead of Q.3. See Appendix 8 for the credibility ranking. Higher rank means higher credibility.

Chapter 5: Discussion and Opportunities for Future Research

Overview

This thesis compiles a series of investigation focusing on anchor-based methodology with a particular aim of developing a systematic, logical approach for selecting an optimal anchor-based MID to enhance the interpretation of PROM results in the context of incorporating patients' perspectives in clinical research and evidence-based decision-making. This concluding chapter discusses the main findings, the strengths and limitations, as well as an exploration of opportunities and directions for future research.

Important Findings

This thesis aimed to address a critical issue for those who use anchor-based MID to aid the interpretation of PROM results: when a number of widely varying MID estimates are available, how to identify, from among the many, the optimal MID? Failure to identify the optimal MID may result in serious misinterpretations of the PROM results though obtained from well-designed trials and meta-analyses.

We began this thesis by comprehensively summarizing the reported methodological issues in estimating anchor based MIDs among the current literature (**Chapter 2**). We used qualitative synthesis to identify items related to selecting anchor-based MIDs. These items listed a spectrum of potential issues important to selecting MIDs, including MID methodology issues related to anchor and PROM, MID statistical issues, generalizability of MID and existing suggestions for MID selection. We did not distinguish suitable from unsuitable items for selecting MIDs in this study. These items, however, provided a conceptual framework to inform the following development of a systematic selection approach for optimal MIDs.

The systematic survey suggested that selecting an MID for use should prioritize MIDs with the greatest credibility. Informed by the systematic survey above, however, to better assess anchor-based MID credibility, the existing credibility instrument needs an extension. Though the instrument has captured most credibility criteria for a trustworthy MID [1], we should also consider the extent to which the PROM and anchor are measuring the same construct –construct proximity, which could also affect the credibility of MIDs. MIDs estimated from very closely related constructs of the anchor and PROM would likely result in a higher correlation than those estimated from anchors with constructs that differ from the PROM. When assessing MID credibility, if investigators did not report the correlation between PROM and anchor, the subjective assessment of construct proximity between PROM and anchor would be a potential alternative. We thus supplemented the existing credibility instrument [1] by adding the construct proximity assessment item as an alternative item for the correlation item (i.e., the correlation between PROM and anchor) and tested the reliability of this extension item, which showed acceptable inter-rater reliability (weighted Cohen's kappa=0.74) (**Chapter 3).**

In **Chapter 4**, we presented the systematic step-by-step approach for selecting an optimal MID. Informed by the systematic survey and expert views, we have iteratively refined the selection approach using real-world examples from the MID inventory [2]. An optimal MID, at least, should be methodologically sound. For the selection, we thus prioritize the methodological rigor. Through credibility assessment[1], we select the most credible MID estimates. If the most credible MIDs fall in a relatively narrow range, investigators choose the median as the optimal MID. Secondarily, an optimal MID should, as far as possible, match the intended application contexts. Evidence

suggesting contextual differences might exist include: the most credible MIDs are inconsistent, or the most credible MIDs are consistent with one another but their median is not near the median of all MIDs. In our approach, either of these findings mandates further exploring contextualized factors to explain the variability among the MIDs. The potential contextualized factors that deserve a consideration could come from the suggestions of previous researchers and include intervention (e.g., surgical vs conservative treatments), patient condition (e.g., knee vs hip osteoarthritis), baseline disease severity, patient age, follow-up duration, socioeconomic status, geography, and sex (**Chapter 2**). When we find evidence to confirm the impacts of contextualized factors on the variability of the MID estimates, to ensure the MID applicable to the contexts of interest, selecting and applying the MID must be context dependent.

Strengths and Limitations

The major strength of this thesis is its innovation, which improves the application of PROs and anchor-based MID. The important advances presented here include the refinement of the existing MID credibility assessment instrument and the development of the selection approach of an optimal MID, which facilitate the appraisal and identification of MID estimates for researchers to enhance the interpretation of PROM results in clinical research and evidence-based decision making. By far, our systematic survey represents the most comprehensive qualitative summary of the methodological issues of anchor-based MID. While no known standards of an optimal MID, through collecting rich information—from both the systematic survey and the expert committee—to inform the development of the selection approach, we first proposed the framework of an optimal MID. That is, being methodologically sound and contextually applicable.

Beyond the limitations noted in the individual chapters, this thesis does not address the development of knowledge translation (KT) and implementation strategies for the selection approach we have created. The potentially relevant stakeholders who will benefit from the selection approach could include anyone who wants to incorporate patients' perspectives or PROs in their research or practice, for example, clinicians, systematic reviewers, guideline developers and policy makers. As a starting point of promoting the application of the selection approach, we have published our work in medical journals, of which the audiences include most of the stakeholders. Future work with research communities focusing on patient reported outcomes could help further identify the barriers and facilitators of the dissemination of the selection approach.

Implications and Opportunities for Future Research

This thesis provided the important considerations to researchers who aim to estimate anchor based MIDs. To obtain a trustworthy estimate, in addition to follow the methodological standards suggested by our systematic survey, researchers should pay further attention to the issues about anchor. We use anchor to reflect the important change of the latent construct of the PROM [3-5]. When the two constructs of the PROM and the anchor are substantially different, though with high correlation, inferring the change on the anchor to the change on the PROM violates intuitive interpretation, challenging the usefulness of the anchor to estimate MIDs for the PROM [6 7]. On the other hand, unequivocally, a modest correlation is the prerequisite of a valid anchor [8 9]. MIDs estimated from very closely related constructs of the anchor and PROM would likely result in a higher correlation than those estimated from anchors with constructs that differ from the PROM. Therefore, the future design of anchor should include the consideration of construct proximity and use anchors with constructs as similar to the constructs of the PROMs as possible.

With the growing emphasis on patient-centered care, using PROMs to directly assess patient's health conditions gradually become a common practice [10-12]. MIDs, as an important benchmark to aid the interpretation of the magnitude of treatment effects for PROMs, constitute a crucial part to assess the benefits received after an intervention from the patients' viewpoint [3 13]. An optimal MID would contribute to more trustworthy interpretation and guide the clinicians, researchers, guideline developers, health policy makers and regulatory authorities to make more informed benefit-risk assessments for interventions. The selection approach proposed here helps the choice of such optimal MID and avoid the risk of choosing MIDs at discretion to fit the research purposes, improving the transparency of health research.

Ideally, after selection, the optimal MID would not change as new evidence emerges and thus become a unique standard to aid the interpretation for a given PROM. We, however, should bear in mind that due to the limited available data of MIDs for certain PROMs, when new estimates emerge, the optimal MIDs may be different. It is, however, possible that the larger the number of high credibility, consistent MIDs, the more compelling the case that a definitive optimal MID has been established; the less impacts of contextualized factors on MID variability, the more likely that the optimal MID would be stable. Currently, the assessment of the optimal MID remains uncertain and open to debate. Grading the certainty of the optimal MID, therefore, could be a potential area for subsequent research. Once we establish the grading standards, we would not only interpret the PROMs based on the optimal MID available, but know the certainty of the 'optimality', leading to more trustworthy interpretation and evidence-based decision-making.

Reference

1. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, Zeraatkar D, Bhatt M, Jin X, Brignardello-Petersen R, Urquhart O, Foroutan F, Schandelmaier S, Pardo-Hernandez H, Vernooij RW, Huang H, Rizwan Y, Siemieniuk R, Lytvyn L, Patrick DL, Ebrahim S, Furukawa T, Nesrallah G, Schünemann HJ, Bhandari M, Thabane L, Guyatt GH. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *Bmj* 2020;369:m1714.
2. Carrasco-Labra A, Devji T, Qasim A, Phillips MR, Wang Y, Johnston BC, Devasenapathy N, Zeraatkar D, Bhatt M, Jin X, Brignardello-Petersen R, Urquhart O, Foroutan F, Schandelmaier S, Pardo-Hernandez H, Hao Q, Wong V, Ye Z, Yao L, Vernooij RWM, Huang H, Zeng L, Rizwan Y, Siemieniuk R, Lytvyn L, Patrick DL, Ebrahim S, Furukawa TA, Nesrallah G, Schünemann HJ, Bhandari M, Thabane L, Guyatt GH. Minimal important difference estimates for patient-reported outcomes: A systematic survey. *J Clin Epidemiol* 2021;133:61-71.
3. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10(4):407-15.
4. McClimans L. Interpretability, validity, and the minimum important difference. *Theor Med Bioeth* 2011;32(6):389-401.
5. Engel L, Beaton DE, Touma Z. Minimal Clinically Important Difference: A Review of Outcome Measure Score Interpretation. *Rheum Dis Clin North Am* 2018;44(2):177-88.
6. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, Guyatt GH. Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. *J Clin Epidemiol* 2009;62(4):374-9.
7. 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.
8. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. *J Clin Epidemiol* 2002;55(9):900-8.
9. Brozek 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.
10. Kornowski R. Patient-reported outcome measures in cardiovascular disease. *European Heart Journal-Quality of Care and Clinical Outcomes* 2023;9(2):119-27.
11. Nelson EC, Eftimovska E, Lind C, Hager A, Wasson JH, Lindblad S. Patient reported outcome measures in practice. *Bmj* 2015;350
12. Lohr KN, Zebrack BJ. Using patient-reported outcomes in clinical practice: challenges and opportunities. *Quality of Life Research* 2009;18:99-107.
13. Schünemann HJ, Akl EA, Guyatt GH. Interpreting the results of patient reported outcome measures in clinical trials: the clinician's perspective. *Health and quality of life outcomes* 2006;4(1):1-8.