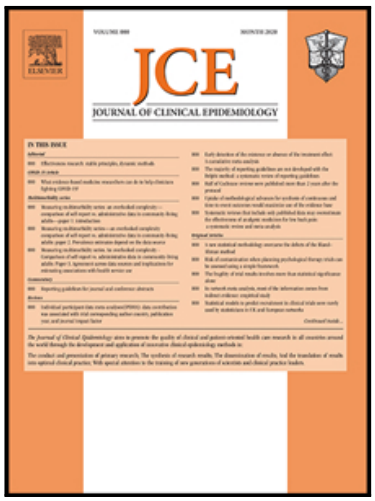
Journal Pre-proof



Minimal important changes in standard deviation units are highly variable and no universally applicable value can be determined

Yasushi Tsujimoto , Tomoko Fujii , Yusuke Tsutsumi , Yuki Kataoka , Aran Tajika , Yohei Okada , Alonso Carrasco-Labra , Tahira Devji , Yuting Wang , Gordon H. Guyatt , Toshi A. Furukawa

PII: S0895-4356(22)00023-3
DOI: https://doi.org/10.1016/j.jclinepi.2022.01.017
Reference: JCE 10753

To appear in: *Journal of Clinical Epidemiology*

Accepted date: 20 January 2022

Please cite this article as: Yasushi Tsujimoto , Tomoko Fujii , Yusuke Tsutsumi , Yuki Kataoka , Aran Tajika , Yohei Okada , Alonso Carrasco-Labra , Tahira Devji , Yuting Wang , Gordon H. Guyatt , Toshi A. Furukawa , Minimal important changes in standard deviation units are highly variable and no universally applicable value can be determined, *Journal of Clinical Epidemiology* (2022), doi: https://doi.org/10.1016/j.jclinepi.2022.01.017

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Inc.

Journal Pre-proof

Minimal important changes in standard deviation units are highly variable and no universally applicable value can be determined

Yasushi Tsujimoto[1,2,3], Tomoko Fujii[1,4], Yusuke Tsutsumi[3,5,6], Yuki Kataoka[3,7,8,9], Aran Tajika[1], Yohei Okada[10], Alonso Carrasco-Labra[11], Tahira Devji[12],Yuting Wang[12], Gordon H Guyatt[12], Toshi A Furukawa[1]

[1]Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine / School of Public Health, Kyoto, Japan
[2]Department of Nephrology and Dialysis, Kyoritsu Hospital, Japan
[3]Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan
[4]Intensive Care Unit, Jikei University Hospital, Tokyo, Japan
[5]Human Health Sciences, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[6]Department of Emergency Medicine, National Hospital Organization Mito Medical Center, Ibaraki, 280 Sakuranosato Ibarakimachi Higashiibarakigun, Ibaraki, Japan
[7]Department of Healthcare Epidemiology, Kyoto University Graduate School of Medicine / School of Public Health, Kyoto, Japan
[8]Department of Internal Medicine, Kyoto Min-Iren Asukai Hospital, Kyoto, Japan
[9]Section of Clinical Epidemiology, Department of Community Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan
[10]Department of Preventive Services, Kyoto University Graduate School of Medicine / School of Public Health, Kyoto, Japan
[11]Center for Integrative Global Oral Health, University of Pennsylvania, School of Dental Medicine, Philadelphia, PA, United States.
[12]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Canada

[*]Corresponding author:

Toshi A. Furukawa

Department of Health Promotion and Human Behavior, School of Public Health in the Graduate School of Medicine, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan

Phone: +81-75-753-9491

Fax: +81-75-753-4641

Email: furukawa@kuhp.kyoto-u.ac.jp

Journal Pre-proof

Journal Pre-proof

Abstract

Objectives

This study aims to describe the distribution of anchor-based minimal important change (MIC) estimates in standard deviation (SD) units and examine if the robustness of such estimates depends on the specific SD used or on the methodological credibility of the anchor-based estimates.

Design and Setting

We included all anchor-based MIC estimates from studies published in MEDLINE and relevant literature databases upto October 2018. Each MIC was converted to SD units using baseline, endpoint, and change from baseline SDs. We performed a descriptive analysis of MICs in SD units and checked how the distribution would change if MICs with low methodological credibility were excluded from the analysis.

Results

We included 1,009 MIC estimates from 182 studies. The medians and interquartile ranges of MICs in SD units were 0.43 (0.25 to 0.69), 0.42 (0.22 to 0.70), and 0.51 (0.28 to 0.78) for baseline, endpoint, and change SD units, respectively. Some MICs were extremely large or small. The distribution did not change significantly after excluding MICs estimated by less credible methods.

Conclusions

The size of the universally applicable MIC in SD units could not be determined. Anchor-based MICs in SD units were widely distributed, with more than half in the range of 0.2 to 0.8.

Keywords

Minimal important change, patient-reported outcome, anchor-based method, distribution-based method, effect size

Journal Pre-proof

Journal Pre-proof

What is new?

Key findings

- Converting the anchor-based minimal important changes (MICs) to standard deviation (SD) units resulted in highly variable, with more than half in the range of 0.2 to 0.8.
- Such wide variation was not changed significantly by excluding MICs estimated by less credible methods.
- Several MICs were extremely large (>10 SD units) or small (zero change).

What this study adds to what was known?

- Previous studies reported a universal value for MIC in SD units, but our results show that it is impossible to determine a constant value due to the wide range of MIC in SD units.
- We found some MICs that have questionably small or large anchor-based MICs when converting them into SD units.
- Most MICs were estimated from small sample size, and the correlation between anchor and patient-reported outcome measures were seldom reported.

What is the implication and what should change now?

- We should stop blindly using MICs in SD units.
- Users should carefully check the methodology to estimate the anchor-based MIC when the MIC in SD units are very small or large.
- To improve the quality of reporting, including correlations between anchors and PROMs, guidelines of reporting items or sample size estimations are warranted.

Journal Pre-proof

Introduction

The minimal important change (MIC), also often called the minimal important difference (MID), for a continuous measure refers to “the smallest difference that patients perceive as beneficial” [1]. MIC

Journal Pre-proof

is used widely and increasingly to interpret the magnitude of changes in patient-reported outcome measures (PROMs). Although the terms MIC and MID are often used interchangeably, it has been proposed that the former be used for longitudinal within-person changes in scores and the latter for cross-sectional between-person differences [2]. In this study we used the term MIC to refer to the former, namely the longitudinal within-person changes in scores.

MIC for a PROM is determined mainly in two ways: an anchor-based method and a distribution-based method. An anchor-based method that uses an external criterion for the PROM is considered the gold standard, rather than a distribution-based method that uses statistical parameters in estimating a MIC [3]. However, anchor-based MICs are not always available. In such a case, we may rely on distribution-based MICs which require a plausible value of MIC in SD unit that predicts the anchor-based MIC. The relationships between the two methods have been debated: several studies reported that MIC estimates by the anchor-based methods agree well with half the standard deviation (SD) of a PROM [4, 5], whereas other studies have shown otherwise [6, 7].

However, these studies did not take the methodological quality of the primary studies or characteristics of PROMs into account. Their samples were more anecdotal than comprehensive. It is also unclear which SD should be used in the distribution-based method, as the baseline SD, the change SD and the endpoint SD of the PROM may be appreciably different from each other. The current study, therefore, aims first to describe the anchor-based MIC estimates in SD units and then to examine if the robustness of such estimates depends on the SD to be used or on the methodological quality of the anchor-based estimates.

Methods

The protocol was published in protocols.io (dx.doi.org/10.17504/protocols.io.bxpqpmmw).

The Minimal Important Difference Inventory

The Minimal Important Difference Inventory dataset included 5,324 MIC estimates derived from 585 studies up to October 2018 [1, 8]. The dataset was derived from a systematic search using MEDLINE, EMBASE, CINAHL and PsycINFO for studies published between 1989 and October 2018, and additional relevant citations from the PROQOLID internal library, relevant reviews and eligible studies.

This dataset included MIC estimates reported in studies estimating anchor-based MICs for PROMs in adolescents (≥13 to 17) or adult (≥18) populations. PROMs of interest included health-related quality of life, functional ability, symptom severity and psychological distress and well-being. It included any MIC irrespective of the participants' condition or disease, type of intervention used in the eligible studies, or nature of the anchor. It excluded systematic reviews of studies examining MICs; conference abstracts; studies in which authors explicitly targeted a moderate or large important difference as opposed to a MIC; a combined anchor- and distribution-based approach; and estimates obtained using pooled data from multiple cohorts (e.g., different primary investigations). More details of the inclusion and exclusion criteria for the Minimal Important Difference Inventory can be found elsewhere [8].

Eligible criteria for the study

In the current study, we included studies that estimated MICs by the mean change methods (i.e., the MIC is the absolute mean change in PROM scores over time within the subgroup of participants who reported they were slightly improved or deteriorated) using a global rating of change as the anchor. We excluded studies or MIC estimates in which the SD of the PROMs were unavailable because the authors failed to report the SD itself, or did not report the standard error (SE), the confidence interval (CI), or interquartile range (IQR) with the number of participants.

Data extraction

Teams of two independent investigators extracted the following variables from the included studies [8]: the country of the study; population demographics; types of a PROM; interventions administered in the context of the MIC estimation; anchor details (i.e., type, constructs, range of options/categories/values, and threshold selected to represent a "small but important change"); a MIC estimate, its associated measure of variability, and direction; the number of patients informing the MIC estimate, and credibility ratings of the MIC estimates [9]. The credibility was rated according to five core criteria: the anchor was rated by the patient; the anchor was interpretable and relevant to the patient; the MID estimate was precise; the correlation between the anchor and the outcome measure reported by the patient was satisfactory (a correlation coefficient of at least 0.5); and the authors select a threshold on the anchor that reflected a small but important difference. We classified types of PROMs in two main categories with two and four subcategories: 1) generic (health profiles and utility measures), and 2) specific (disease/condition-specific, symptom-specific, function-specific, and population-specific) according to the previous taxonomy [10]. PROMs categorized into health profiles are instruments that attempt to measure all important aspects of health-related quality of life (HRQOL). Utility measures are derived from economic and decision theory that reflect the preferences of patients for treatment process and outcome. Specific measures focus on aspects of health status that are specific to the area of primary interest. Reviewers, working in pairs, independently conducted the data extraction, resolving disagreements by discussion with input from a third reviewer for quality assurance.

We also extracted the means and the SDs of PROMs in all participants from the primary studies. We classified the SDs into the following categories: the SD of the baseline scores (baseline SD), the SD of endpoint scores (endpoint SD), and the SD of change from baseline scores (change SD). If authors reported other variability measures such as SE, CI, IQR, or range but not SD, we calculated SD using the following formulae.

$\mathrm{SD} = \mathrm{SE} \times \sqrt{N}$.

$\mathrm{SD} = \sqrt{N} \times (\text{upper limit of CI} - \text{lower limit of CI}) / 3.92$.

SD = IQR / 1.35 [11].

SD = Range / 2 [11].

If authors reported the SDs separately in subgroups of the participants (e.g., improved, no change, or deteriorated), we calculated the overall SD using the following formula in the Cochrane handbook [11]. When there were more than two groups to combine, we applied the following formula

Journal Pre-proof

sequentially. However, if the authors reported the variability in subgroups using 95% CI, IQR, or range, we did not calculate the overall SD, because the formulae do not provide consistent values when the number of participants is small [12].

$$SD = \sqrt{\frac{(N_1 - 1)SD_1^2 + (N_2 - 1)SD_2^2 + \frac{N_1 N_2}{(N_1 + N_2)(M_1^2 + M_2^2 - 2M_1 M_2)}}{N_1 + N_2 - 1}}$$

Statistical analysis

Our primary outcome was the distribution of MICs in SD units. We presented the histogram of MICs for three types of SDs; baseline SD, endpoint SD, and change SD. We calculated their means, SDs, 95% CIs, medians, IQRs, and ranges. MICs in SD units were calculated using the following formula.

$$MIC\ in\ SD\ units = \frac{MIC}{SD}$$

To examine if the MICs in SD units depends on the credibility of the estimating methods, we further performed a sensitivity analysis excluding MIC estimates with low credibility. We used the following core criteria to assess the credibility of MIC estimation [9].

1) Is the patient or necessary proxy responding directly to both the PROM and the anchor? – "No" or "Impossible to tell" for low credibility,
2) Is the anchor easily understandable and relevant for patients or a necessary proxy? – "Definitely no", "Not so much" or "Impossible to tell" for low credibility,
3) Has the anchor shown a good correlation with the PROM? - "Definitely no" for low credibility (i.e., a correlation coefficients of less than 0.3)
4) Is the MIC precise? - "Definitely no" for low credibility (i.e., 95% CI range is wider than 50% of MIC estimate or sample size is less than 100)
5) Does the threshold or difference between groups on the anchor used to estimate the MIC reflect a small but important difference? - "Definitely no", "Not so much" or "Impossible to tell" for low credibility

As a previous study found that only a minority of studies estimating MIC provided data on the correlation and precision, we did not exclude those with a judgment of "Not so much" or "Impossible to tell" for these two items [13]. As a post-hoc sensitivity analysis, we presented the distribution of MIC estimates in SD unit, dividing the direction of MIC into improvement, deterioration, and both. We used Stata/SE, V.14.0 (StataCorp, College Station, TexasX, USA) for all analyses.

Results

Figure 1 shows the flow diagram of the present study. Of 14,540 records identified by the search, 585 studies estimated one or more MICs using the anchor-based method. We further excluded studies that did not use the mean change method (n = 346) and studies that did not report variability of PROMs to calculate the SDs (n = 57). Finally, we included 182 studies consisting of 187 PROMs and 1,009 MIC estimates. Table 1 summarizes the characteristics of included studies and PROMs. Most studies recruited adults or the elderly. Of 187 PROMs, 140 (74%) were disease-specific PROMs. Table 2

Journal Pre-proof

Journal Pre-proof

shows the characteristics of MIC estimates in the present study. The correlation between the anchor and PROM was not reported in 591 (59%) and 332 (33%) estimates did not show a good correlation for. Since the number of participants contributing to estimate MICs was small (median 35, IQR 19 to 61), 909 (90%) estimates were not precise (i.e., sample size of less than 100).

Among 1,009 MIC estimates, those in baseline, endpoint, and change SD units were available for 931 (93%), 582 (58%), and 530 (53%), respectively. Figure 2 shows the distribution of MIC estimates in each SD unit. Table 3 describes the summary statistics of MIC estimates in each SD unit. The medians and IQRs of MIC in each SD unit were 0.43 (0.25 to 0.69) for baseline SD units, 0.42 (0.22 to 0.70) for endpoint SD units, and 0.51 (0.28 to 0.78) for change SD units, respectively. The proportion of MIC in each SD unit less than 0.2 or greater than 0.8 was 38%, 42%, and 39% for baseline, endpoint, and change SD units, respectively.

Two studies by the same authors reported eight MICs which had extremely high estimates in baseline SD unit (greater than 10 SD) (Figure 2A). One study recruited patients with trigeminal neuralgia suffering severe pain that interfered greatly with their daily lives and examined the MICs of pain in Visual Analog Scale (VAS) when the patients received percutaneous stereotactic radiofrequency lesioning interventions [14, 15]. Consequently, the baseline SDs were small (e.g., 0.50 for VAS) and the MICs were large (e.g., 8.20 for VAS). This resulted in extremely large MICs in SD units (e.g., 16.4 for VAS). We also found four MICs of 0 [16-19]. Details of such extremely large or small MICs were described in Supplementary Table 1.

A total of 320, 190, and 155 MIC estimates in baseline, endpoint, and change SD units were included in the sensitivity analysis with highly credible estimates. The medians and IQRs of MIC in each SD unit were 0.35 (0.18 to 0.52) for baseline SD units, 0.37 (0.19 to 0.60) for endpoint SD units, and 0.46 (0.26 to 0.69) for change SD units (Table 3). Supplementary Figure 1 shows the distribution of MICs in each SD unit excluding MIC estimates with low credibility. Supplementary Table 2 describes the distribution of MIC estimates in SD unit, dividing the direction of MIC into improvement, deterioration, and both. MICs in SD unit for deterioration had slightly small value and narrow distribution as compared with other directions of MIC.

Discussion

We presented the distribution of anchor-based MIC estimates in SD units using a comprehensive dataset. Neither SDs from baseline score, endpoint score, nor change from baseline score yielded a universally applicable and widely generalizable value of MIC in SD units. The broad distribution of MICs in SD units was not narrowed down by the sensitivity analysis excluding MICs estimated with a less credible methodology. We also found some studies that reported extremely large or small MICs when converted to SD units.

Contrary to 0.5 SD for the MIC reported by Norman et al., there was no constant or consistent value for MIC in SD units that could be universally used for PROMs no matter which SD was used [5]. The wide variation of MICs in SD units might be due to the variations of participants' characteristics at baseline such as disease severity. Substantial heterogeneity of the sample in different studies yielded different SDs and hence different MICs in SD units even when the same absolute MICs were reported. However, a previous study found a large variation in the absolute anchor-based MIC estimates by the same method across studies and across different methods within studies [21]. They concluded it was not clear whether the variation was due to differences between populations or

to conceptual and methodological problems of the MIC methods. We hypothesised that the variation depended on different types of SDs; i.e., baseline SD would be smaller and more heterogeneous than endpoint SD or change SD. However, such a pattern was not observed. We also explored the influence of the methodological quality to estimate the anchor-based MICs on the distribution. However, the sensitivity analysis, which excluded less credible MIC estimates, did not narrow the distribution enough to be universally applicable. Although the post-hoc sensitivity analysis showed a narrower distribution for deterioration as compared with other directions, the width of the IQRs were still large.

We found that approximately 60% of MICs in SD units were within the range from 0.2 to 0.8 (Figure 2). It may be possible to consider a range of 0.2 to 0.8 as a rough guide when an anchor-based MIC for a PROM is missing. If MICs in SD units did not fall in this range, it is questionable whether such very small or large values represent true MICs. In such case, we may rely on the distribution-based approach to interpret the change in a PROM

Several included studies reported extremely large or small MIC estimates in SD units [14-19]. A previous study suggested that estimates of the MICs are substantially larger for subgroups of patients with high baseline values [22]. If the anchor and PROM are not correlated well, the MIC can be zero or grossly erroneous. For example, one may feel slightly worse in a global assessment even if a patient-reported outcome does not change as a result of the treatment expectations. The credibility instrument includes assessment of the correlation between anchor and PROM [9].

We also found much room for improvement in the methodology and reporting in future MIC studies. More than half of the included studies did not report the correlation between the PROM and anchor. The correlation is closely related to the usefulness of anchor-based approaches; an anchor that has low or no correlation with the patient reported outcome measure will likely give inaccurate MID estimates [9]. Also, sample size calculation should be necessary to estimate a precise MIC estimation; only 10% of the included MICs had precise estimates in our sample [9]. The inferences about the magnitude of a treatment effect would differ at the extremes of the CI around of the MIC estimate when the CI is wide. There is a need for reporting guidelines that recommends the assessment of the methodological credibility for MIC estimates.

This study has several limitations. First, as above, we could not exclude the possibility that the wide variation of MICs in SD units was due to the poor methodological credibility of MIC estimates. Our sensitivity analysis, which aimed to exclude the studies with less methodological credibility to estimate the MIC, actually included studies which did not report correlations or studies with small sample sizes. Thus, the analysis was not sufficient to check the robustness of our findings. Second, we did not assess whether the other distribution approach, such as the standard error of the mean (SEM) and the responsiveness statistics, are useful [21]. Third, both the anchor-based method and the distribution-based method have large variability in how they estimate and calculate MICs. In this study we tried to examine their relationships while limiting the former to those using the more rigorous methodology and using different SDs in calculating the latter approach. Finally, our MIC inventory data set was comprehensive only through October 2018. The process of searching for studies published since 2018 October is ongoing and part of a plan to continually add new MICs to the living web-based MID inventory (PROMID, www.promid.org) [8].

Despite the above limitations, this study added new insights to the existing controversy regarding the relationship between SDs and MICs. We systematically retrieved all available studies on anchor-based MICs using rigorous methodology. As a result, this study used the largest and the

Journal Pre-proof

most comprehensive data to assess the relationship between SD and MIC. We differentiated baseline, endpoint, and change SDs, which are often confused.

Conclusion

Converting the anchor-based MICs to SD units resulted in highly variable estimates, which made it difficult to determine a universal value of MICs in SD units for the distribution method. Such variation was unlikely to be changed by types of SDs or the methodological credibility. Thus, there appears no solid basis to use a certain value for MIC in SD units. However, due to the poor reporting of correlations and the small sample sizes in the available studies, there remains a possibility that the wide variation of MICs in SD units was caused by the inadequate methodological quality of the currently available studies of MICs.

Author statement

YTsuj had full access to all data used in the study and takes responsibility for both the integrity of the data and the accuracy of the data analysis. YTsuj, TF, YTsut, YK, AT, ACL, TD, GHG, and TAF developed the study's concept and design. YTsuj, TF, YTsut, YK, AT, OY, ACL, TD, YW, GHG, and TAF acquired the data. YTsuj and TAF analyzed and interpreted the data. YTsuj and TAF drafted the manuscript. TF, YTsut, YK, AT, OY, ACL, TD, YW and GHG critically revised the manuscript for important intellectual content. All the authors provided their final approval of the version submitted for publication and agreed that they were accountable for all aspects of this study.

Journal Pre-proof

Acknowledgements

We would like to thank the team of the minimal important difference inventory study for constructing the dataset (Drs Anila Qasim, Mark R. Phillips, Yuting Wang, Bradley C. Johnston, Niveditha Devasenapathy, Dena Zeraatkar, Meha Bhatt, Xuejing Jin, Romina Brignardello-Petersen, Olivia Urquhart, Farid Foroutan, Stefan Schandelmaier, Hector Pardo-Hernandez, Qiukui Hao, Vanessa Wong, Zhikang Ye, Liam Yao, Robin W.M. Vernooij, Hsiaomin Huang, Linan Zeng, Yamna Rizwan, Reed Siemieniuk, Lyubov Lytvyn, Donald L. Patrick, Shanil Ebrahim, Gihad Nesrallah, Holger J. Schünemann, Mohit Bhandari, and Lehana Thabane.)

We also thank the members of the Research Group on Meta-Epidemiology at the Kyoto University School of Public Health (Drs Akira Onishi, Ethan Sahker, Yan Luo, and Kenji Omae.) They provided many valuable comments on this research.

Footnotes

Contributors

YTsuj had full access to all data used in the study and takes responsibility for both the integrity of the data and the accuracy of the data analysis. YTsuj, TF, YTsut, YK, AT, ACL, TD, GHG, and TAF

developed the study's concept and design. YTsuj, TF, YTsut, YK, AT, OY, ACL, TD, YW, GHG, and TAF acquired the data. YTsuj and TAF analyzed and interpreted the data. YTsuj and TAF drafted the manuscript. TF, YTsut, YK, AT, OY, ACL, TD, YW and GHG critically revised the manuscript for important intellectual content. All the authors provided their final approval of the version submitted for publication and agreed that they were accountable for all aspects of this study.

Funding

The Minimal Important Difference Inventory project was funded in part by the Canadian Institutes of Health Research (CIHR), Knowledge Synthesis grant number DC0190SR. The present study was also funded by Kyoto University School of Public Health Crossover Award.

Role of the funding source

The funding organization did not influence the design or conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; nor the decision to submit this manuscript for publication.

Competing interests

Dr Furukawa reports grants and personal fees from Mitsubishi-Tanabe, personal fees from MSD, grants and personal fees from Shionogi, and personal fees from SONY outside the submitted work. In addition, Dr Furukawa has a patent (2020-548587) concerning smartphone Cognitive Behavior Therapy applications pending and intellectual properties for Kokoro-app licensed to Mitsubishi-Tanabe. The other authors have no conflicts of interest to declare.

Ethical approval

Not required.

Data sharing

The Minimal Important Difference Inventory is the copyright of McMaster University (Copyright © 2018, McMaster University, Hamilton, Ontario, Canada). The Minimal Important Difference Inventory has 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.

Journal Pre-proof

References:

1. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Controlled clinical trials. 1989;10(4):407-15.

2. de Vet HCW, Ostelo RWJG, Terwee CB, van der Roer N, Knol DL, Beckerman H, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2007;16(1):131-42.

3. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. Journal of Clinical Epidemiology. 2003;56(5):395-407.

4. Sloan JA, Loprinzi CL, Kuross SA, Miser AW, O'Fallon JR, Mahoney MR, et al. Randomized comparison of four tools measuring overall quality of life in patients with advanced cancer. J Clin Oncol. 1998;16(11):3662-73.

5. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003;41(5):582-92.

6. Testa MA. Interpreting quality-of-life clinical trial data for use in the clinical practice of antihypertensive therapy. J Hypertens Suppl. 1987;5(1):S9-13.

7. Feinstein AR. Indexes of contrast and quantitative significance for comparisons of two groups. Stat Med. 1999;18(19):2557-81.

8. Carrasco-Labra A, Devji T, Qasim A, Phillips MR, Wang Y, Johnston BC, et al. Minimal important difference estimates for patient-reported outcomes: A systematic survey. J Clin Epidemiol. 2020;133:61-71.

9. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, 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.

10. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med. 1993;118(8):622-9.

11. Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0

: Cochrane; 2019 [updated July 2019. Available from: www.training.cochrane.org/handbook.

.

12. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol. 2005;5:13.

13. 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.

14. Reddy VK, Parker SL, Lockney DT, Patrawala SA, Su PF, Mericle RA. Percutaneous stereotactic radiofrequency lesioning for trigeminal neuralgia: determination of minimum clinically important difference in pain improvement for patient-reported outcomes. Neurosurgery.

Journal Pre-proof

2014;74(3):262-6; discussion 6.

15. Reddy VK, Parker SL, Patrawala SA, Lockney DT, Su PF, Mericle RA. Microvascular decompression for classic trigeminal neuralgia: determination of minimum clinically important difference in pain improvement for patient reported outcomes. Neurosurgery. 2013;72(5):749-54; discussion 54.

16. Jones GL, Radley SC, Lumb J, Farkas A. Responsiveness of the electronic Personal Assessment Questionnaire-Pelvic Floor (ePAQ-PF). Int Urogynecol J Pelvic Floor Dysfunct. 2009;20(5):557-64.

17. Khanna D, Furst DE, Maranian P, Seibold JR, Impens A, Mayes MD, et al. Minimally important differences of the UCLA Scleroderma Clinical Trial Consortium Gastrointestinal Tract Instrument. J Rheumatol. 2011;38(9):1920-4.

18. Leung YY, Zhu TY, Tam LS, Kun EW, Li EK. Minimal important difference and responsiveness to change of the SF-36 in patients with psoriatic arthritis receiving tumor necrosis factor-α blockers. J Rheumatol. 2011;38(9):2077-9.

19. Quittner AL, O'Donnell AE, Salathe MA, Lewis SA, Li X, Montgomery AB, et al. Quality of Life Questionnaire-Bronchiectasis: final psychometric analyses and determination of minimal important difference scores. Thorax. 2015;70(1):12-20.

20. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989;27(3 Suppl):S178-89.

21. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, et al. Mind the MIC: large variation among populations and methods. J Clin Epidemiol. 2010;63(5):524-34.

22. Ward MM, Guthrie LC, Alba M. Dependence of the minimal clinically important improvement on the baseline value is a consequence of floor and ceiling effects and not different expectations by patients. J Clin Epidemiol. 2014;67(6):689-96.

Journal Pre-proof

Figure titles and legends

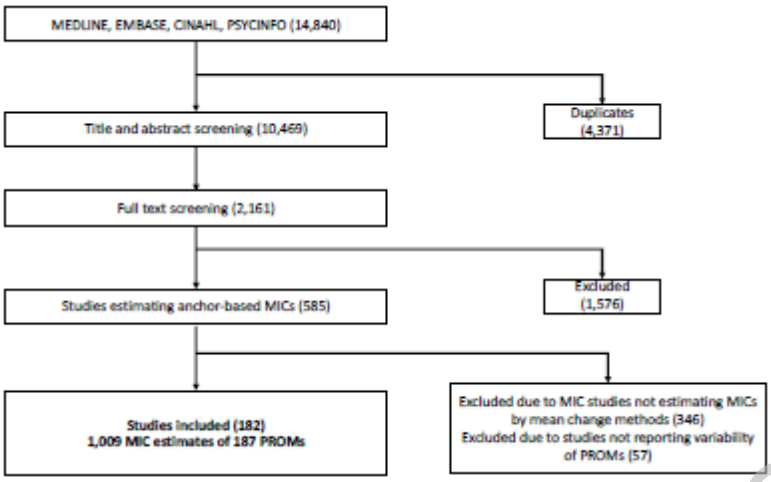


Figure 1. Flow diagram of study eligibility.
Abbreviations: PROM, patient-reported outcome measure; MIC, minimal important change

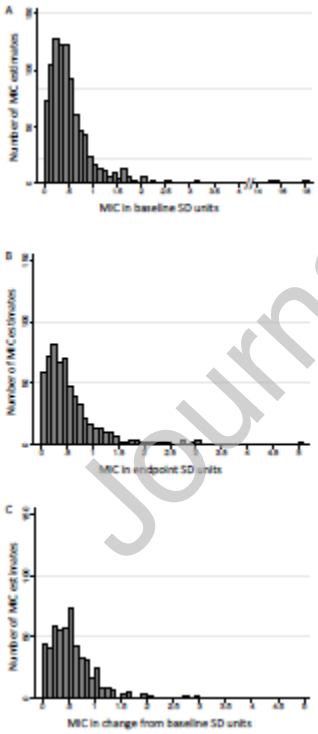


Figure 2. Distribution of minimal important change estimates in each standard deviation unit.
Figure A, B, and C display MICs in baseline, endpoint, and change SD units, respectively.
Abbreviations: MIC, minimal important change; SD, standard deviation

Journal Pre-proof

Journal Pre-proof

Table 1. Characteristics of included studies and patient-reported measures

Characteristics	
Study-level (n = 182)	
Year of publication	2013 (2009 to 2016)
Number of total participants	191 (100 to 306)
Age groups	
Children or adolescents	5 (3)
Adults or elderly	169 (93)
Not reported	8 (4)
Interventions	
Pharmacological intervention	31 (17)
Surgical or invasive intervention	40 (22)
Rehabilitation	17 (9)
Mixture	50 (27)
Others	20 (11)
Not available	24 (13)
PROM-level (n = 187)	
Types of PROMs*	
Generic, Health profile	7 (4)
Generic, Utility measure	7 (4)
Specific, Disease/condition	140 (74)
Specific, Symptom	26 (14)
Specific, Function	6 (3)

Journal Pre-proof

Specific, Population	1 (1)
Number of MIC estimates per PROM	2 (1 to 4)

Notes: Values are expressed as number (percentage) or median (interquartile range). *According to the previous taxonomy of types of PROMs (reference [9]). Utility measures are derived from economic and decision theory that reflect the preferences of patients for treatment process and outcome. Specific measures focus on aspects of health status that are specific to the area of primary interest. Abbreviations: PROM, patient-reported outcome measure; MIC, minimal important change.

Journal Pre-proof

Journal Pre-proof

Table 2. Characteristics of minimal important difference estimates

Characteristics	
Number of participants for MIC estimation	35 (19 to 61)
Direction of MIC	
Improvement	651 (65)
Deterioration	218 (22)
Any change	110 (11)
Unclear	30 (3)
Credibility assessment of MIC estimates*	
Is the patient or necessary proxy responding directly to both the PROM and the anchor?	
Yes	984 (98)
No / impossible to tell	25 (2)
Is the anchor easily understandable and relevant for patients or necessary proxy?	
Definitely yes / to a great extent	974 (97)
Definitely no / not so much / impossible to tell	35 (3)
Has the anchor shown good correlation with the PROM?	
Definitely yes / to a great extent	86 (9)
Definitely no / not so much / impossible to tell	332 (33)
Not reported	591 (59)
Is the MIC precise?	
Definitely yes / to a great extent	100 (10)
Definitely no / not so much / impossible to tell	909 (90)

Journal Pre-proof

Does the threshold or difference between groups on the anchor used to estimate the MIC reflect a small but important difference?	
Definitely yes / to a great extent	671 (67)
Definitely no / not so much / impossible to tell	3385 (33)

Notes: Values and Abbreviations; MIC, minimal important change. * According to the credibility instrument for the MIC estimates (reference [8]). Abbreviations: MIC, minimal important difference; PROM, patient-reported outcome measure

Table 3. Summary statistics of minimal important change estimates in each SD unit

Type sf SD units	**Mean (SD)**	**95%CI**	**Median**	**IQR**	**Range**
Primary analysis					
Baseline (n = 931)	0.67 (1.52)	0.58 to 0.77	0.43	0.25 to 0.69	0 to 18.11
Endpoint (n = 582)	0.55 (0.52)	0.50 to 0.59	0.42	0.22 to 0.70	0 to 5.02
Change (n = 530)	0.59 (0.47)	0.55 to 0.63	0.51	0.28 to 0.78	0 to 3.42
Sensitivity analysis*					
Baseline (n = 320)	0.42 (0.37)	0.38 to 0.46	0.35	0.18 to 0.52	0 to 2.77
Endpoint (n = 190)	0.43 (0.32)	0.38 to 0.47	0.37	0.19 to 0.60	0 to 2.32
Change (n = 155)	0.50 (0.32)	0.45 to 0.55	0.46	0.26 to 0.69	0 to 1.72

*Sensitivity analysis excluding minimal important change estimates with less credible methodology. Abbreviations: SD, standard deviation; CI, confidence interval; IQR, interquartile range

Journal Pre-proof