Use of the minimal important difference as a criterion for clinical importance – are we off track?

Stefan Moosmayer, MD PhD

PII: S2666-6391(22)00101-8
DOI: https://doi.org/10.1016/j.xrrt.2022.09.007
Reference: XRRT 188

To appear in: JSES Reviews, Reports, and Techniques

Received Date: 6 April 2022
Revised Date: 18 September 2022
Accepted Date: 23 September 2022

Please cite this article as: Moosmayer S, Use of the minimal important difference as a criterion for clinical importance – are we off track?, JSES Reviews, Reports, and Techniques (2022), doi: https://doi.org/10.1016/j.xrrt.2022.09.007.

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 The Author(s). Published by Elsevier Inc. on behalf of American Shoulder & Elbow Surgeons.
Use of the minimal important difference as a criterion for clinical importance – are we off track?

Running title: Minimal important difference and clinical importance

Author:
Stefan Moosmayer, MD PhD
Department of Orthopedic Surgery
Martina Hansens Hospital
Doenskiveien 8, 1346 Gjettum, Norway
st.moos@online.no

Institutional review board or ethical committee approval was not required for this review article.

Disclaimers:
Funding: No funding was disclosed by the authors.
Conflicts of interest: The authors, their immediate families, and any research foundation with which they are affiliated have not received any financial payments or other benefits from any commercial entity related to the subject of this article.
Use of the minimal important difference as a criterion for clinical importance – are we off track?

Level of evidence: Narrative Review

Keywords: Minimal important difference; clinical importance; shoulder treatment

Over the last 30 years, we have seen a paradigm shift in the evaluation of shoulder treatment results from the clinician’s perspective to the patient’s perspective. Patient-reported outcome measures (PROMs) are now widely used as the primary outcome in clinical trials of shoulder patients. A major issue with the use of PROMs, however, is that whether a change of score represents clinical relevance for the patient or not is not possible to ascertain directly from the scores. For instance, a mean improvement of 8 points on a PROM ranging from 0 to 100 in a clinical study may represent statistical significance, but it is difficult to say if it represents a relevant improvement for the patient. To bridge the gap between statistical significance and clinical importance, the concept of the minimal clinically important difference (MCID) was introduced by Jaeschke and was later renamed the minimal important difference (MID). According to Jaeschke, the MID represents the smallest change in an outcome score that patients might perceive as an important improvement (or deterioration) of their health problem and can be used as a threshold for clinical relevance in research and clinical practice. The MID concept implies that a treatment result must at least exceed the PROM’s MID value to represent relevance for the patients. Over the last few years, MID values have been determined for many of the PROMs commonly used in shoulder research. Shoulder surgeons and researchers have welcomed the values, as they seem to give a simple solution to the complex problem of determining clinical importance. It has now become common practice to judge the clinical relevance of a treatment result by comparing it to one of the MID values that have been reported for the study PROM. However, as intuitively as the MID principle may appear, it has important conceptual limitations that affect the interpretability of the MID.
Moreover, many of the studies that report MID values for commonly used shoulder scores have been criticized for poor study methodology and poor reporting of their results. Many of the published MID values have been found to be of questionable credibility and may be inappropriate as reference points for the clinical relevance of treatment effects. Researchers and readers of scientific reports should be aware of the weaknesses and limitations of the MID to avoid misinterpretations and inadequate conclusions. In the following report, we want to inform readers of some of the most critical aspects of the MID concept.

First, MID results are affected by the methodological approach, selected for calculation. Different methods of calculation, including distribution-based and anchor-based approaches, are in use and result in a wide range of MID values for the same outcome measure. Distribution-based approaches are based on statistical characteristics of the study sample, such as the 0.5 standard deviation (SD) or the standard error of the mean (SEM) of the study score, and do not account for the patients’ perspective. Today, distribution-based estimates are considered to be of second value and should only be applied in the absence of anchor-based MID estimates. In anchor-based approaches, the results from the PROM are linked to a level of improvement selected by the patients on a global rating of change (GRC) scale (e.g., compared to baseline, my shoulder is “worse”, “unchanged”, “slightly better”, “much better”, and “very much better”). The MID can then be determined by one of three methods, either as the score change from baseline to follow-up in the “slightly better” group, as the difference in score change between the “unchanged” and the “slightly better” group, or by a receiver operating characteristic (ROC) curve. As the results from these different approaches may vary considerably, it is difficult to define one generally valid MID level for a given PROM. This is well exemplified by the results from five recent studies reporting MID estimates for the American Shoulder and Elbow Surgeons (ASES) score for patients treated surgically for a rotator cuff tear. Table I shows eleven different values from different studies and calculation
methods for the MID ranging from 6.1 to 27.1 points. Values vary not only between studies, but also within studies, which confirms the influence of the calculation method on the results. In the absence of evidence regarding whether one method of calculation is superior to the other, the importance of one single MID value remains unclear. Today’s frequent practice of selectively picking one among several published MID values and presenting it as a kind of score constant seems inappropriate. Second, the definition of the threshold for a small but important change on an anchor will influence the MID results. On a GRC scale with the response options “a little better”, “much better” and “very much better”, it may be agreed that the “a little better” group represents a better indicator for a change that is small but important than the “much better” or “very much better” groups. Some studies, however, use GRC scales with up to 7 response categories for improvement (and corresponding categories for deterioration), such as “almost the same”, “a little better”, “somewhat better”, “moderately better”, “a good deal better”, “a great deal better”, and “a very great deal better”, and there is no evidence or consensus regarding which group (or combination of groups) should be selected to represent a small but important improvement. The investigators must decide subjectively where to place the cut off on the GRC scale, and the decision will directly influence the resulting MID value. Another common but equally questionable anchor-based approach was used by Kukkonen et al to determinate a MID value for the Constant score for patients treated by arthroscopic rotator cuff repair. The study patients were offered only two response options on the GRC scale: “Is your shoulder better or worse?”, and the MID was determined as the mean score change in the “better” group. Such analyses may lead to inflated MID estimates, as the “better” group most certainly not only includes patients who would have rated their outcome as “a little better” but also all those who would have selected “much better” or “very much better” on a rating scale with these options. A closer analysis of reported MID values has
shown that many of them are based on inappropriate anchor thresholds. In an instrument development and reliability study performed on 3389 anchor-based MID estimates for 358 PROMs across all clinical areas, the authors found that only 43% of the anchor thresholds reflected a small but important difference. MID values therefore must be given together with specific information about the anchor that was used for determination. Only then is it possible to assess whether the given MID truly represents a small but important difference.

Third, MID values are often presented as a single point but should be given with a measure of precision, such as the 95% confidence interval. MID values depend on the heterogeneity of the studied population and thus carry a degree of uncertainty. Presenting a MID as a single cutoff point for clinical importance gives the false impression that a treatment result just below the MID does not represent clinical relevance, while a slightly higher result, just exceeding the MID, does. The presentation of results as a range, which likely includes the true value in the population, is essential in most clinical research, and thus it is surprising that MID results are often given and interpreted without such information. According to a recently published review of studies determining MID values for shoulder outcomes, one of three studies failed to provide critical information about the data distribution. Two other reports have shown that many MID values, even those based on large study samples, are imprecise and contain large confidence intervals, reflecting the great variation in people’s perception of a relevant improvement. Examples for large CIs around MID values which have been determined in studies of four commonly used shoulder scores are given in table II. MID values, which are reported without an adequate measure of precision, lack interpretability and have little credibility as a threshold for clinical relevance on PROM scales.

A second important but seldomly reported measure for the credibility of an anchor-derived MID is the degree of correlation between the rating from the GRC scale and the results from the PROM. When patients are asked to answer the anchor question, they must be able to
compare their present shoulder status to their status at baseline. Recalling earlier health states, however, is difficult, especially at longer follow-ups (recall bias), and patients may base their answer more on their current condition than on the degree to which they have changed. To confirm that the rating on the GRC reflects true change rather than present state, it should be shown that there is a correlation between the rating on the GRC and the baseline score and that the rating on the GRC is more closely correlated with the change in score than to the score at follow-up. Furthermore, a correlation of approximately 0.5 should be demonstrated between the rating on the GRC scale and the change in the PROM to support that the GRC is adequately linked with the measured score difference. It has been claimed that the correlations between the anchor and the PROM represent the single most important aspect of the credibility of an MID. It is therefore a cause of serious concern that the majority of studies of MID values for shoulder PROMs fail to report such correlations.

Fourth, many authors seem to overlook the fact that MID values depend on the characteristics of the patient group from which they have been calculated and may not be valid for a patient group with different characteristics. Consequently, a universally valid MID for a given PROM does not exist. MID results have been shown to be particularly sensitive to differences in the baseline score. It appears plausible that the value of a clinically important change differs between highly symptomatic patients and patients with fewer symptoms. Other characteristics that may influence MID values are age, sex and treatment methods. Uncritically selecting an MID value among values reported for a study’s PROM is therefore not acceptable. Authors must show that the selected value was determined based on a group of subjects with comparable baseline characteristics and which were treated by the same treatment method. Especially questionable is the widespread use of one single MID value as a benchmark for clinical importance in review articles. Such use would only be acceptable if all studies that are included in the review were performed on comparable patient groups.
Finally, MID values are usually determined based on a longitudinal analysis as a treatment-related score difference from baseline to follow-up within a patient group and not as a between-group difference. The validity of the common practice of assessing clinical relevance by comparing the result from a between-group comparison directly to an MID value which was derived from a within-group analysis is therefore debatable. The recommended procedure for the use of the MID in this case is to use the MID as the cut off value in a proportion analysis where patients who meet the MID are classified as responders and where the proportion of responders are compared between groups. Using the MID value as a threshold for responders and nonresponders, however, may not be very informative in the comparison of highly effective treatment measures, as are common in shoulder surgery. Most results in both groups will exceed the MID, which is too low to differentiate the groups, and existing between-group differences may remain undetected. Alternative threshold values, such as the Patient Acceptable Symptom State (PASS) or the Substantial Clinical Benefit (SCB) should be used in this situation.

Conclusions

The assessment of the clinical importance of a treatment outcome is important in both research and clinical practice. In recent years, the MID has been introduced as a simple and easily accessible measure for clinical relevance. A closer look at reported MID values and their application, however, shows that there are important problems that affect their interpretability. It has been demonstrated that MID values are sensitive to their method of calculation and to the specific characteristics of the anchor that was used for their determination. The resulting wide range of MID values for the same shoulder PROM makes it difficult to know which of the values, if any, represents a valid threshold of clinical importance. Today’s frequent practice of drawing firm conclusions about the clinical
relevance of treatment results by comparing them uncritically to one of the published MID
values of the study PROM seems inappropriate. The use of the MID as a metric for clinical
relevance should be accompanied by a discussion of the validity of the selected values and
their suitability for the specific study population. Researchers and clinicians should be careful
not to overinterpret the MID to avoid overlooking important results from well-designed
studies.

References:


Table I: Reported MID values for the ASES score for surgically treated patients with a rotator cuff tear.

<table>
<thead>
<tr>
<th>Author, year</th>
<th>Method of calculation</th>
<th>MID results (points)</th>
<th>Measure of variation</th>
<th>No. of patients</th>
<th>F.-up (yrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gagnier, 2018*</td>
<td>Anchor-based (10-item question): Difference in score change method</td>
<td>21.9</td>
<td>95% CI: 4.1 to 39.6</td>
<td>213</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Distribution based: 0.5 SD</td>
<td>26.9</td>
<td>Not given</td>
<td>194</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Distribution based: 0.33 SD</td>
<td>17.9</td>
<td>Not given</td>
<td>194</td>
<td></td>
</tr>
<tr>
<td>Cvetanovich, 2019</td>
<td>Anchor based (15-item question): ROC curve derived</td>
<td>11.1</td>
<td>Not given</td>
<td>288</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Distribution based: 0.5 SD</td>
<td>11.7</td>
<td>Not given</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gowd, 2019*</td>
<td>Anchor based (15-item question): ROC curve derived</td>
<td>11.1</td>
<td>Not given</td>
<td>89</td>
<td>1</td>
</tr>
<tr>
<td>Tashjian, 2020</td>
<td>Anchor based (4-item question): Difference in score change method</td>
<td>27.1</td>
<td>Not given</td>
<td>202</td>
<td>1</td>
</tr>
<tr>
<td>Kim, 2020</td>
<td>Anchor based (4-item question): ROC curve derived</td>
<td>21.0</td>
<td>Not given</td>
<td>82</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Anchor based (4-item question): Difference in score change method</td>
<td>14.0</td>
<td>Not given</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malavolta, 2022</td>
<td>Anchor based (2-item question): ROC curve derived</td>
<td>6.1</td>
<td>Not given</td>
<td>289</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Distribution based: 0.5 SD</td>
<td>10.5</td>
<td>Not given</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimum detectable change approach</td>
<td>26.3</td>
<td>Not given</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MID, minimal important difference; ASES, American Shoulder and Elbow Surgeons; CI, confidence interval; SD, standard deviation; ROC, receiver operating characteristic.

* In this study 57% of patients were surgically treated, and 43% were conservatively treated.
Table II: Reported MID values for outcome scores of the shoulder together with their confidence intervals.

<table>
<thead>
<tr>
<th>Author, year</th>
<th>Score</th>
<th>Method of calculation</th>
<th>MID result (points)</th>
<th>95% CI</th>
<th>No. of patients†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Werner, 2016²²</td>
<td>ASES</td>
<td>Difference in score change method</td>
<td>13.5</td>
<td>4.8 to 22.3</td>
<td>490/67</td>
</tr>
<tr>
<td>Gagnier, 2018⁵</td>
<td>ASES</td>
<td>Mean change method</td>
<td>21.9</td>
<td>4.1 to 39.6</td>
<td>213/19</td>
</tr>
<tr>
<td>Gagnier, 2018⁵</td>
<td>WORC</td>
<td>Mean change method</td>
<td>282.6</td>
<td>-39.0 to 604.3</td>
<td>213/19</td>
</tr>
<tr>
<td>Zhou, 2018²³</td>
<td>SANE</td>
<td>Mean change method</td>
<td>27.3</td>
<td>16.7 to 38.3*</td>
<td>222/21</td>
</tr>
<tr>
<td>Kanto, 2021¹²</td>
<td>Constant</td>
<td>Difference in score change method</td>
<td>24</td>
<td>18 to 31</td>
<td>179/NG</td>
</tr>
</tbody>
</table>

ASES, American Shoulder and Elbow Surgeons; WORC, Western Ontario Rotator Cuff Index; SANE, single assessment numeric evaluation; MID, minimal important difference; CI, confidence interval; NG, Not given.

*Given as the 90% CI.
†The first value is the total number of patients in the study. The second value is the number of patients reporting a level of improvement on the global rating of change scale which was selected by the authors to represent a minimal important difference.
Conflict of interest statement:

Stefan Moosmayer

This author, his immediate family, and any research foundation with which he is affiliated did not receive any financial payments or other benefits from any commercial entity related to the subject of this article.

No outside funding or grant was received for this article

Date: 6.3.22  Signature: [Signature]

Stefan Moosmayer
Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: