Mathematical coupling may account for the association between baseline severity and minimally important difference values

John Patrick Browne*, Jan H. van der Meulen, James D. Lewsey, Donna L. Lamping, Nick Black

Health Services Research Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, UK
Accepted 23 October 2009

Abstract

Objective: To generate anchor-based values for the “minimally important difference” (MID) for a number of commonly used patient-reported outcome (PRO) measures and to examine whether these values could be applied across the continuum of preoperative patient severity.

Study Design and Setting: Six prospective cohort studies of patients undergoing elective surgery at hospitals in England and Wales. Patients completed questionnaires about their health and health-related quality of life before and after surgery. MID values were calculated using the mean change score for a reference group of patients who reported they were “a little better” after surgery minus the mean change score for those who said they were “about the same.” Pearson’s correlation was used to examine the association between baseline severity and change scores in the reference group. Baseline severity was expressed in two ways: first in terms of preoperative scores and second in terms of the average of pre- and postoperative scores (Oldham’s method).

Results: Of the 10 PRO measures examined, eight demonstrated a moderate or high positive association between preoperative scores and MID values. Only two measures demonstrated such an association when Oldham’s measure of baseline severity was used.

Conclusion: In general, there is little association between baseline severity and MID values. However, a moderate association persists for some measures, and it is recommended that researchers continue to test for this relationship when generating anchor-based MID values from change scores.

Keywords: Quality of life; Surgery; Psychometrics; Longitudinal studies; Health services research; Outcome assessment

1. Introduction

Interpretability is a key challenge for researchers interested in measures of health status and health-related quality of life. Patient-reported outcome (PRO) measures do not produce intuitively meaningful data, and this makes it difficult to interpret the meaning of differences between and within groups and individuals. The meaning of unit changes in the measures is unclear, because the metrics change from instrument to instrument and also because of unfamiliarity with their use, unlike, for example, measures of blood pressure [1]. It is also important to ensure that statistically significant results have clinical or social significance [2]. The importance of interpretability was emphasized in guidance issued in 2006 by the United States Food and Drug Administration, which recommended the specification of a “minimally important difference” (MID) when developing PRO measures. The MID has been defined as “the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side-effects and excessive cost, a change in the patient’s management” [3]. MIDs may be derived using either distribution- or anchor-based methods.

1.1. Distribution-based approaches

Distribution-based approaches use statistical aspects of representative samples to determine an MID. The most commonly used methods describe change in terms of standard deviation (SD) units. For example, Cohen’s effect size formula defines an MID as 0.20 for “small” effects, 0.50 for “moderate” effects, and 0.80 for “large” effects [4]. The effect size effect size suffers from sample dependence: The greater the variability within a sample, the higher the SD and the higher the MID. The standard error of the measurement method has been suggested to produce a distribution-based MID that is relatively constant when measured
What is new?

- We have provided anchor-based MID values for the first time for a number of commonly used generic and disease-specific patient-reported outcome measures.
- Previous research on the relationship between baseline severity and anchor-based MID values may have reached inappropriate conclusions because of flawed statistical analyses.
- When an unbiased test is used there is, in general, a low positive association between baseline severity and anchor-based MID values and for some measures there is a low negative association.

in different samples of patients [5]. The main disadvantage of both methods is the arbitrariness of the qualitative thresholds and the lack of a known relationship to patient experience. The remainder of this article, therefore, focuses on an alternative approach, in which MID values are “anchored” to a known categorical change in health status.

1.2. Anchor-based approaches: between-group comparisons

Two types of anchor-based MID can be derived: between group and within group. Between-group values are based on a comparison of scores of patients in different clinical groups. For example, Deyo et al. compared sickness impact profile (SIP) scores in rheumatoid arthritis patients classified into one of four severity categories as defined by the American Rheumatism Association. Differences in average scores between adjacent categories were then used as MID values for the SIP [6]. A similar approach was used by Kulkarni to produce an MID for the Hydrocephalus Outcome Questionnaire. In this case, the average scores for children rated by clinicians as “not at all impaired” were compared with those rated as “very mildly impaired” [7]. A disadvantage of this approach is that clinical severity and patient-perceived health status, although correlated to some extent, are different constructs. Clinical severity, therefore, can only shed limited light on the meaning behind a patient’s response. Furthermore, it is unclear if cross-sectional MIDs can be applied in longitudinal studies of health status.

1.3. Anchor-based approaches: within-group comparisons

Within-group MIDs can be derived by calculating the mean change score in a reference group of patients deemed to have experienced minimally important change. The reference group may be defined according to external clinical criteria or criteria based on the patient’s own perspective. In an example of the clinically driven approach, Eton et al. used the Eastern Cooperative Oncology Group Performance Status Rating scale to identify three types of patients: those who improved, those who were stable, and those who deteriorated over time. MID values for the Functional Assessment of Cancer Therapy—Lung Symptom Index-12 (FACT-LSI12) were then produced by calculating mean change scores on the FACT-LSI12 for each of the three groups [8]. The use of MID reference groups defined according to clinical measures has been criticized for prioritizing criteria that have no known relationship to the patient experience [9]. Revicki et al. recommended that “the patient’s perspective be given the most weight since these are patient reported outcome, although the clinician’s perspective is considered important as well” [10].

The first use of a patient-referenced approach to MID generation for a quality-of-life instrument was reported by Jaeschke et al. in 1989 [3]. Patients were asked to rate their symptom change on a 15-point Transition Rating Index (TRI) ranging from −7 to +7, where zero represents no change. Those who answered from +3 to −3 on this scale were identified as the reference group, and their mean change score on the quality-of-life instrument was used as the MID. Juniper et al. used the same TRI but included only those patients who scored between +1 and −1 in the reference group [1]. In both the Jaeschke et al. and Juniper et al. studies, patients who reported improvement and deterioration were grouped together. It is now commonly accepted that this should be avoided, because the mean absolute magnitude of change scores in improving and deteriorating patients is not the same [11]. Separate MIDs for patients that report improvement and deterioration have been derived in more recent studies [12–14]. A further modification of the Jaeschke et al. approach involves recalibrating change scores in patients who report minimal change to take into account change scores for patients who report no change in their symptoms. For example, Coyne et al. [15] produced an MID for a quality-of-life measure in patients with an overactive bladder by subtracting the mean change score for patients reporting no benefit from the mean change scores for patients reporting little benefit. This recalibration is performed, because any deviation from zero in the change scores of patients who report no change in symptoms is considered to represent “measurement error.”

1.4. Minimally important difference values for commonly used patient-reported outcome measures

The Department of Health in England has signaled an intention to move toward the routine use of PRO measures to assess the benefits of health care [16]. The Department commissioned our group to perform a systematic review to identify the most appropriate measures to use for five high-volume elective procedures: cataract surgery, groin hernia
repair, varicose vein surgery, hip replacement, and knee replacement. We found that procedure-specific MID values had not been produced for the preferred measures [17]. A value for the MID is also unavailable for the Sino-Nasal Outcome Test (SNOT, 22-item version), a measure deemed the best available for sinonasal surgery patients in a recent review [18]. In this study, we generated anchor-based MID values for seven different disease-specific and generic PRO measures that are commonly used to assess the benefits of high-volume elective surgery.

1.5. Influence of baseline severity

A number of studies suggest that MIDs based on change scores lack stability across the continuum of baseline severity. Specifically, previous studies report that MID values are higher in patients with greater baseline severity and lower in patients with lower severity. The studies cover a diverse set of patients, including those with back pain [19–21], isolated trauma of the extremities [22], obesity [23], conditions requiring elective surgery [24], osteoarthritis [25], localized musculoskeletal pain [26], conditions requiring emergency care [27], and chronic pain [28]. This evidence has been cited in a number of review articles, and it is now conventional wisdom that MID values are associated with baseline severity [11,29,30]. The results are usually explained in psychophysical terms: patients with a high baseline severity need a higher magnitude of change to perceive a clinically meaningful change in their condition. To deal with this phenomenon, it has been suggested that PRO measures should have separate MID values for discrete categories of baseline severity [11,23,25,27,28]. As a result, it is unclear to what extent the influence of baseline severity on anchor-based MIDs is a clinical phenomenon or a statistical artifact. To address this question, an unbiased test of the relationship between baseline severity and change is required. Such a test was proposed by Oldham in 1962. Rather than test the correlation between \( x \) and \( x - y \), he proposed testing the correlation between \((x + y)/2\) and \(x - y\). This correlation equals zero for two series of independent random numbers \(x\) and \(y\) with the same SD and, thus, avoids the statistical problems described earlier. Why should we correlate change with an average of pre- and posttreatment scores when we are interested in the relationship between pretreatment scores and change? It can be shown that this correlation is a test of the differences in the variances between an initial measurement and a repeated measurement and that Oldham’s coefficient will equal zero if there is no difference in the variances [31]. Oldham reasoned that if the effectiveness of a treatment is related to baseline severity, we should observe a “shrinking” in the posttreatment variance compared with the pretreatment variance. This will occur, because the proportional response to treatment will cause posttreatment scores to converge around the mean. This can be demonstrated using the hypothetical example of three patients completing a health status measure that is bounded by 0 and 100 and which gives lower scores for patients with greater impairment. Patient 1 has a score of 20 before surgery, patient 2 scores 40, and patient 3 scores 60. Because improvement is greater for patients with higher baseline severity in this example, patient 1 improves by 40 points to 60, patient 2 by 30 points to 70, and patient 3 by 20 points to 80. As a result, the dispersion of scores as measured by the SD around the mean is halved from 20 before surgery to 10 after surgery. Although Oldham’s method cannot specify the incremental change produced by increases in baseline severity, it does give an unbiased test of correlation for differential baseline effects.

1.6. Statistical flaws in previous studies

Because anchor-based MIDs are simply mean change scores in a subsample of patients, studies that investigate the relationship between such MIDs and baseline severity must deal with the same statistical challenges faced by any enquiry into the relationship between change and initial value. Mathematical coupling occurs “when one variable directly or indirectly contains the whole or part of another” [31]. Because change scores are the pretreatment score minus the posttreatment score, they contain the pretreatment score. Mathematical coupling can lead to an artificially inflated association between initial value and change score when correlation or regression is used. This was demonstrated by Oldham in 1962, who showed that, for two series of independent random numbers \(x\) and \(y\) with the same SD, a strong correlation \((r = 0.71)\) is observed between \(x\) and \(x - y\) [32]. Using the same random numbers, it can also be shown that mean values for \(x - y\) vary across different strata for \(x\). For example, when \(x\) and \(y\) are bounded by 0 and 100, the mean value of \(x - y\) for the lowest quartile of \(x\) is approximately \(-37\), and the mean value of \(x - y\) for the highest quartile of \(x\) is approximately \(+37\).

Previous research on the relationship between MIDs derived from change scores and baseline severity has not adequately accounted for mathematical coupling. The studies have used correlations [24], regression methods [26], and the comparison of mean change scores for different baseline strata [19–23,25,27,28]. As a result, it is unclear to what extent the influence of baseline severity on anchor-based MIDs is a clinical phenomenon or a statistical artifact. To address this question, an unbiased test of the relationship between baseline severity and change is required. Such a test was proposed by Oldham in 1962. Rather than test the correlation between \(x\) and \(x - y\), he proposed testing the correlation between \((x + y)/2\) and \(x - y\). This correlation equals zero for two series of independent random numbers \(x\) and \(y\) with the same SD and, thus, avoids the statistical problems described earlier. Why should we correlate change with an average of pre- and posttreatment scores when we are interested in the relationship between pretreatment scores and change? It can be shown that this correlation is a test of the differences in the variances between an initial measurement and a repeated measurement and that Oldham’s coefficient will equal zero if there is no difference in the variances [31]. Oldham reasoned that if the effectiveness of a treatment is related to baseline severity, we should observe a “shrinking” in the posttreatment variance compared with the pretreatment variance. This will occur, because the proportional response to treatment will cause posttreatment scores to converge around the mean. This can be demonstrated using the hypothetical example of three patients completing a health status measure that is bounded by 0 and 100 and which gives lower scores for patients with greater impairment. Patient 1 has a score of 20 before surgery, patient 2 scores 40, and patient 3 scores 60. Because improvement is greater for patients with higher baseline severity in this example, patient 1 improves by 40 points to 60, patient 2 by 30 points to 70, and patient 3 by 20 points to 80. As a result, the dispersion of scores as measured by the SD around the mean is halved from 20 before surgery to 10 after surgery. Although Oldham’s method cannot specify the incremental change produced by increases in baseline severity, it does give an unbiased test of correlation for differential baseline effects.

1.7. Study aims

Our aim was to generate anchor-based MID values for a number of commonly used PRO measures in patients undergoing elective surgery and then use Oldham’s method to examine whether these values were related to preoperative severity.
2. Methods

Patients were from six cohorts undergoing elective operations: sinonasal surgery, cataract surgery, groin hernia repair, varicose vein surgery, hip replacement, and knee replacement. All patients were treated at hospitals in England and Wales, and the research protocols were approved by all relevant institutional ethics committees. Consecutive patients aged 16 years or more were invited to complete questionnaires about their health and quality of life before and after surgery. Patients judged to be incapable of completing a written questionnaire in English (because of cognitive impairment, poor sight, literacy, or language comprehension problems) were not asked to participate. Postoperative questionnaires were mailed to the patients’ home address 3 months after surgery (6 months for those undergoing hip or knee replacement).

The sinonasal surgery cohort consisted of patients undergoing nasal polypectomy and/or surgery to alleviate chronic rhinosinusitis at 87 hospitals from April 2000 to May 2001. A complete description of the methods used in this study has been published elsewhere [33]. Both the pre- and postoperative patient questionnaires included a 22-item version of the SNOT, a previously validated measure of health-related quality of life in sinonasal disease [34]. The SNOT-22 total score ranges from 0 to 110, and higher scores represent greater impairment.

The cataract surgery cohort consisted of patients undergoing cataract removal at 11 hospitals from January 2006 to June 2007. Both the pre- and postoperative questionnaires contained the Visual Function-14 Index (VF-14) of functional impairment in patients with cataract [35]. The VF-14 total score ranges from 0 to 100, and higher scores indicate better visual function. The EuroQol (EQ-5D), a generic measure of [36] was also administered but was found to be unresponsive to change in this cohort; hence, results are not reported.

The groin hernia repair cohort consisted of patients undergoing inguinal or femoral hernia repair at 18 hospitals from January to September 2006. Both the pre- and postoperative questionnaires contained the EQ-5D. As no diseasespecific measure for hernia repair has been validated, the physical component summary (PCS) of the UK version of the Short Form (SF)-36 was also used [37]. The EQ-5D produces values ranging from −0.549 to 1.0, and the PCS is scored on a 0–100 scale. Higher scores on both measures indicate lower impairment.

The varicose vein surgery cohort consisted of patients undergoing surgery for varicose veins at 16 hospitals from January to November 2006. Patients completed both the EQ-5D and the Aberdeen Varicose Veins Questionnaire (AVVQ) [38] before and after surgery. The AVVQ is scored on a 0–60 scale, and higher scores indicate greater impairment.

Both the hip and knee replacement cohorts consisted of patients undergoing unilateral hip or knee replacement or resurfacing at 14 hospitals from January 2006 to February 2007. Patients completed the EQ-5D both before and after surgery. In addition, the hip replacement cohort completed the Oxford hip score [39], and the knee replacement cohort completed the Oxford knee score [40]. Both questionnaires are scored on a 12–60 scale, and higher scores indicate greater impairment.

2.1. Generating minimally important difference values

In addition to the measures previously described, all postoperative questionnaires contained a TRI comparing pre- and postoperative health status on a 5-point scale (1 = much better, 2 = a little better, 3 = about the same, 4 = a little worse, and 5 = much worse). The MID for each of the seven PRO measures described earlier was defined as the mean change score for those who reported they were “a little better” minus the mean change score for those who were “about the same.” As the scoring direction for the measures varied, we calculated change scores for all instruments such that positive MIDs would indicate improvement.

2.2. Validity of Transition Rating Index

To check the construct validity of the TRI, we calculated the Spearman rank-order correlation between TRI responses and change scores for each measure. We calculated change scores for all PRO measures such that higher positive Spearman coefficients would indicate higher levels of validity. When judging the magnitude of such correlations, Guyatt et al. proposed a standard of 0.5 [41]. This was chosen on the basis that “We believe that correlations less than 0.5 provide grounds for doubting the validity of that particular transition rating.” However, less stringent standards are often used. For example, “Cohen’s rule” [4]—where \( r < 0.3 \) is considered a low correlation, \( 0.3–0.6 \) moderate, and greater than 0.6 high—is often cited to provide guidelines for interpreting the magnitude of correlations in construct-validity studies. As both the Guyatt and Cohen thresholds are arbitrary, they are used only to guide the interpretation of findings in this article and not to disqualify a particular MID value.

2.3. Association between minimally important differences and baseline severity

Pearson’s correlation was used to examine the association between baseline severity and change scores in the subsample of patients who reported they were “a little better” after surgery. Baseline severity was expressed in two ways: in terms of preoperative scores and the average of pre- and postoperative scores (Oldham’s method). When calculating Pearson correlations, scores on the VF-14, EQ-5D, and SF-36 PCS were reversed so that positive coefficients would indicate a positive association between baseline severity and change scores across all measures. A direct comparison of the correlated pre- and postoperative variances was also performed. The formula for this test is:
\[ t = \frac{(s_x^2 - s_y^2)\sqrt{n - 2}}{2s_x s_y \sqrt{1 - r_{xy}^2}} \]

where \( s_x \) is the preoperative variance, \( s_y \) is the postoperative variance, and \( r_{xy} \) is the correlation between pre- and postoperative scores. A \( P \)-value based on the \( t \) distribution was derived using \( n - 2 \) degrees of freedom [42]. \( P \)-Values lower than 0.05 were considered a statistically significant result, and all correlation coefficients were assumed to be equal to zero in the null hypothesis. Stata software (Stata Corp., College Station, TX, USA) was used for all calculations [43].

3. Results

The number of patients in each cohort who consented to participate and completed a baseline questionnaire were as follows: 2,561 sinonasal surgery, 866 cataract surgery, 570 groin hernia repair, 363 varicose vein surgery, 512 hip replacement; and 526 knee replacement. The baseline characteristics of patients are shown in Table 1.

The number of patients in each cohort who returned a postoperative questionnaire was as follows: 2,141 sinonasal surgery patients (response rate, 84%); 750 cataract surgery patients (87%); 441 groin hernia repair patients (77%); 269 varicose vein surgery patients (74%); 445 hip replacement patients (87%); and 461 knee replacement patients (88%). For all six cohorts, there were no statistically significant differences in preoperative disease-specific measures between those who did or did not return a postoperative questionnaire.

Table 2 shows preoperative, postoperative, and change scores for each disease-specific PRO measure broken down according to TRI responses. The sample sizes in Table 2 refer to patients who successfully completed both the pre- and postoperative versions of the measure in question and a TRI and are, therefore, slightly smaller than the sample sizes for all patients who returned postoperative questionnaires. Table 3 uses the same approach to display data for generic measures. For all six cohorts, the baseline characteristics of patients who were included in Tables 2 and 3 were similar to the baseline characteristics of patients who were not included. Responses on the TRI were positively skewed, with around 90% of patients in all cohorts reporting they were “much better” or “a little better” after surgery. As a result, the sample sizes for patients reporting no improvement or deterioration were quite small.

3.1. Validity of Transition Rating Index

The TRI generally discriminated between patients with different levels of pre- to postoperative change. Unexpected discrepancies in the magnitude of change scores across certain TRI categories observed for some measures (e.g., Oxford hip score) are likely because of the small number of patients in these categories. Responses to the TRI were moderately correlated with change scores for the following measures: SNOT-22, 0.49; Oxford hip score, 0.40; Oxford knee score, 0.56; SF-36 PCS in hernia patients, 0.30; EQ-5D in varicose vein patients, 0.33; and EQ-5D in knee replacement patients, 0.42. Low correlations (<0.3) were observed for the following measures: VF-14, 0.19; AVVQ, 0.22; EQ-5D in hernia patients, 0.24; EQ-5D in hip replacement patients, 0.22.

3.2. Minimally important difference values

The MID values produced for each measure are as follows: SNOT-22, 8.9; AVVQ, 2.4; Oxford hip score, 8.4; Oxford knee score, 3.8; SF-36 PCS in hernia patients, 3.8; EQ-5D in hernia patients, 0.03; EQ-5D in varicose vein patients, 0.04; EQ-5D in hip replacement patients, 0.13; and EQ-5D in knee replacement patients, 0.15. Our approach for generating MIDs failed to produce a meaningful value for the VF-14. The mean VF-14 change score in patients who reported that they were “a little better” was 5.5 compared with 7.3 for the patients who reported they were “about the same” (Table 2). By subtracting the latter from the former, we produced a negative MID value of −1.8, indicating that patients perceived a degree of improvement in their condition although their VF-14 scores had deteriorated.

3.3. Association between minimally important differences and baseline severity

In the subgroup of patients who report they are “a little better” after surgery, there is, for most measures, a strong
positive association between the severity of preoperative scores and the magnitude of change scores (Table 4). The Pearson correlation coefficient is statistically significant in all but two cases: the SF-36 PCS in hernia repair patients and the Oxford hip score. This association is diminished in all samples and for all measures when baseline severity is expressed as an average of pre- and postoperative scores using Oldham’s method. In six instances, the Pearson correlation switches from moderate to low in magnitude when Oldham’s method is used: SNOT-22 (0.51 vs. 0.13); VF-14 (0.51 vs. 0.03); EQ-5D in hip replacement patients (0.68 vs. 0.30). In the final two instances, there was a switch from a low positive correlation to a negative correlation: SF-36 PCS in hernia patients (0.20 vs. −0.15) and the Oxford hip score (0.16 vs. −0.33). In both these instances, the postoperative SD is greater than the preoperative SD (Table 3), indicating that variance expanded rather than shrunken after surgery. In general, the SDs observed for pre- and postoperative scores are very similar (Tables 2 and 3). The P-values derived from the variance ratio test closely match those derived from Oldham’s correlation coefficient. A statistically significant difference between pre- and postoperative variance was reduced from high to moderate in magnitude: EQ-5D in varicose vein patients (0.75 vs. 0.31) and EQ-5D in knee replacement patients (0.68 vs. 0.30). In the final two instances, there was a switch from a low positive correlation to a negative correlation: SF-36 PCS in hernia patients (0.20 vs. −0.15) and the Oxford hip score (0.16 vs. −0.33). In both these instances, the postoperative SD is greater than the preoperative SD (Table 3), indicating that variance expanded rather than shrunken after surgery. In general, the SDs observed for pre- and postoperative scores are very similar (Tables 2 and 3). The P-values derived from the variance ratio test closely match those derived from Oldham’s correlation coefficient. A statistically significant difference between pre- and postoperative variance was

### Table 2

<table>
<thead>
<tr>
<th>Measure</th>
<th>Much better</th>
<th>A little better</th>
<th>About the same</th>
<th>A little worse</th>
<th>Much worse</th>
<th>All patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preop. SNOT-22</td>
<td>1,271</td>
<td>40.3 (19.0)</td>
<td>448 42.9 (20.0)</td>
<td>246 44.4 (20.8)</td>
<td>47 48.9 (22.7)</td>
<td>31 59.7 (23.9)</td>
</tr>
<tr>
<td>Postop. SNOT-22</td>
<td>1,271</td>
<td>17.0 (15.0)</td>
<td>448 33.4 (18.1)</td>
<td>246 43.8 (21.8)</td>
<td>47 54.7 (20.4)</td>
<td>31 67.7 (21.8)</td>
</tr>
<tr>
<td>SNOT-22 change</td>
<td>1,271</td>
<td>23.3 (18.4)</td>
<td>448 9.5 (15.9)</td>
<td>246 0.6 (15.3)</td>
<td>47 −5.7 (15.2)</td>
<td>31 −8.0 (20.3)</td>
</tr>
<tr>
<td>Preop. VF-14</td>
<td>605</td>
<td>84.0 (16.3)</td>
<td>85 81.8 (17.7)</td>
<td>23 77.0 (19.6)</td>
<td>12 78.6 (16.2)</td>
<td>13 74.3 (27.5)</td>
</tr>
<tr>
<td>Postop. VF-14</td>
<td>605</td>
<td>96.2 (8.5)</td>
<td>85 87.3 (17.2)</td>
<td>23 84.4 (19.3)</td>
<td>12 74.9 (13.2)</td>
<td>13 55.3 (36.3)</td>
</tr>
<tr>
<td>Preop. AVVQ</td>
<td>187</td>
<td>15.4 (7.5)</td>
<td>53 18.8 (7.9)</td>
<td>16 20.8 (10.0)</td>
<td>6 17.3 (7.7)</td>
<td>3 16.1 (5.0)</td>
</tr>
<tr>
<td>Postop. AVVQ</td>
<td>187</td>
<td>18.8 (7.5)</td>
<td>53 13.8 (8.3)</td>
<td>16 18.2 (12.6)</td>
<td>6 15.3 (5.5)</td>
<td>3 20.5 (7.0)</td>
</tr>
<tr>
<td>AVVQ change</td>
<td>187</td>
<td>17.6 (7.0)</td>
<td>53 5.0 (6.9)</td>
<td>16 2.6 (11.2)</td>
<td>6 2.0 (9.0)</td>
<td>3 −4.4 (11.7)</td>
</tr>
<tr>
<td>Preop. Oxford hip score</td>
<td>380</td>
<td>42.6 (8.0)</td>
<td>34 46.2 (6.6)</td>
<td>11 40.4 (8.6)</td>
<td>5 44.2 (6.6)</td>
<td>6 44.8 (2.8)</td>
</tr>
<tr>
<td>Postop. Oxford hip score</td>
<td>380</td>
<td>20.8 (7.7)</td>
<td>34 32.5 (6.6)</td>
<td>11 35.2 (8.9)</td>
<td>5 34.6 (3.6)</td>
<td>6 51.2 (3.1)</td>
</tr>
<tr>
<td>Oxford hip score change</td>
<td>380</td>
<td>21.9 (8.7)</td>
<td>34 13.6 (6.7)</td>
<td>11 5.2 (7.0)</td>
<td>5 9.6 (5.4)</td>
<td>6 −6.3 (5.0)</td>
</tr>
<tr>
<td>Preop. Oxford knee score</td>
<td>325</td>
<td>41.4 (7.4)</td>
<td>72 41.4 (9.1)</td>
<td>29 39.7 (6.8)</td>
<td>15 37.8 (6.4)</td>
<td>12 41.7 (10.4)</td>
</tr>
<tr>
<td>Postop. Oxford knee score</td>
<td>325</td>
<td>22.9 (7.2)</td>
<td>72 31.4 (8.0)</td>
<td>29 33.4 (6.7)</td>
<td>15 38.3 (7.0)</td>
<td>12 46.7 (7.2)</td>
</tr>
<tr>
<td>Oxford knee score change</td>
<td>325</td>
<td>18.5 (8.2)</td>
<td>72 10.0 (7.4)</td>
<td>29 6.2 (6.9)</td>
<td>15 −0.5 (5.7)</td>
<td>12 −5.1 (8.5)</td>
</tr>
</tbody>
</table>

**Abbreviations:** TRI, Transition Rating Index; SNOT, Sino-Nasal Outcome Test; SD, standard deviation; Preop., preoperative; Postop., postoperative; AVVQ, Aberdeen varicose veins questionnaire.

* Patients who completed both the pre- and postoperative measure and the TRI.
observed only for the SNOT-22 (preoperative SD = 20.0; postoperative SD = 18.1; \( P = 0.001 \)) and for the EQ-5D in knee replacement patients (preoperative SD = 0.32; postoperative SD = 0.25; \( P = 0.01 \)).

### 4. Discussion

#### 4.1. Minimally important difference values

We have provided anchor-based MID values for the first time for a number of commonly used generic and disease-specific PRO measures. The TRI method for obtaining anchor-based MID values failed with one measure (VF-14) because of a discrepancy between the VF-14 change scores observed for patients who reported they were “about the same” after surgery and those who reported they were “a little better.” The TRI method had only moderate construct validity in terms of correlation with prospectively obtained change scores, and in four measures (VF-14, AVVQ, EQ-5D in hernia, and EQ-5D in hip replacement), the correlation was classified as “low.” The use of TRIs has been criticized for insufficient attention to validity [44], and our results indicate that a strong association between TRIs and change scores cannot be assumed. The association is particularly weak for generic measures.

It is possible to compare the MID values produced in this study for the EQ-5D with a previous study that used similar methods [45]. The previous research produced MID values that ranged from 0.002 to 0.26 and supports the finding of the current study that anchor-based MID values for the EQ-5D are different in different patient populations. This is a major challenge for a generic measure, such as the EQ-5D, as change is supposed to have the same meaning across all populations [46]. It also seems unfair to set different MID thresholds for different treatments when using a generic measure: Why should knee replacement require a threshold of 0.15 on the EQ-5D that is five times larger than the threshold for hernia surgery (0.03)? It can be argued, however, that the different MIDs produced for the EQ-5D are simply another artifact of mathematical coupling. With respect to the subsamples that say they are “a little better” after surgery, patients undergoing knee replacement (mean preoperative EQ-5D score = 0.35) represent a far more severe strata of baseline severity than those undergoing hernia surgery (mean preoperative EQ-5D = 0.79). As discussed in the Introduction, a substantial difference in change scores can be expected for patients from different baseline strata based on random numbers alone. To counteract this effect, it would be more sensible to calculate an average MID value for the EQ-5D from a range of patient groups representing the spectrum of baseline severity rather than produce separate values for each group. This approach would also be more equitable when comparing the effectiveness of different treatments.

#### 4.2. Impact of baseline severity on minimally important difference values

This study demonstrates that previous research on the relationship between baseline severity and anchor-based MID values may have reached inappropriate conclusions because of flawed statistical analyses. Oldham’s method demonstrates that, in general, there is a low positive association between baseline severity and MID values, and for some measures, there is a low negative association. A related test of the similarity between pre- and postoperative variances showed that, in general, there was little evidence that postoperative scores converge toward the mean after surgery as
would be expected if change was related to initial score. It is not possible to specify an exact threshold for the correlation statistic at which one can reject the hypothesis that baseline severity is associated with MID values. However, the guidelines on the magnitude of correlations produced by Cohen ($r = 0.3$), as outlined in the Methods section of this article, may prove to be a useful guide for interpretation. We found that 8 out of 10 correlations exceeded this threshold when we examined the association between preoperative and change scores in the reference group of patients who said they were “a little better” after surgery. In contrast, in only two instances (Oxford knee score and EQ-5D with varicose vein patients) did the correlation between change scores and the Oldham measure of baseline severity reach the Cohen threshold for a moderate positive correlation (>0.30). A moderate negative correlation (−0.33) was produced for the Oxford hip score. The $P$-values observed for all measures are very similar for Oldham’s method and the variance ratio test. These methods use the same mathematical assumptions [47]; hence, this is to be expected. The power to detect a statistically significant difference varied between subsamples, and it is of note that the variance ratio test produced a statistically significant result in two groups with relatively large sample sizes. Of greater interest is the magnitude of coefficients, as these indicate the extent to which Oldham’s method changes the size of effect. For example, although we observed a statistically significant difference between the pre- and postoperative variance of the SNOT-22 score ($P = 0.001$), our interpretation should be guided more by the low correlation (0.13) produced by Oldham’s method. These findings produce strong evidence that a single value for the MID is appropriate for many PRO measures. It is important to note that the use of multiple MID values creates logistical problems for researchers hoping to use PRO measures. One of the most important uses of MID values is in sample size calculations. If we have multiple MID values, then we must also have multiple sample sizes—a situation to be avoided if at all possible. If MIDs are to be used in sample size calculations, then it is reasonable to use one value if we are confident that any association between MID values and baseline severity is a statistical artifact because of mathematical coupling. It is important to recognize that, in some cases, a real association between pretreatment states and the response to treatment will apply. Within psychophysics, this phenomenon has been described as “the law of initial value” and was first described by Wilder in 1958 [48]. Within clinical research, a recent reanalysis of studies of the association between baseline pocket probing depth and the effectiveness of guided tissue regeneration in periodontal disease found that a real association persisted in 3 out of 12 studies even after adjustment using Oldham’s method [47]. It is recommended, therefore, that researchers continue to test for a possible association between baseline severity and MID values using the methods described in this study.

4.3. Study weaknesses

Although the overall sample sizes in this study were quite large, the number of patients who stated they were “a little better” or “about the same” after surgery was relatively small in some cohorts. This decreased the precision of the results reported for certain measures. There is a particular danger that some of the MID values reported in this study lack sufficient precision to be used without future validation. It is striking that the MID produced for the Oxford hip score (8.4) is more than twice the value for the Oxford knee score (3.8). These measures use the same metric and address very similar patient groups; hence, the result is surprising. It is likely that the very small sample sizes for the reference group of patients who report they are “a little better” ($n = 34$) or “about the same” ($n = 11$) after hip replacement has contributed to this result. It may also be argued that the 5-point scale we used was not sufficiently discriminatory in selecting patients who had experienced a minimally important improvement after surgery, although there is no published evidence to support one TRI method over another in this regard.

5. Conclusions

This study generated anchor-based MID values for seven PRO measures that are commonly used with patients undergoing elective surgery. In four instances (VF-14, AVVQ, EQ-5D in hernia, and EQ-5D in hip replacement), there was poor evidence for the validity of the method used, and the MID values produced for these measures should be treated with appropriate caution. For all measures other than the SNOT-22, it is recommended that larger studies be performed to validate the MID values produced. MID values generally show no association with baseline severity when Oldham’s method is used. A moderate positive association is observed for the EQ-5D in varicose vein surgery and knee replacement, and a moderate negative association is observed for the Oxford hip score. This implies that, for most of the PRO measures examined, anchor-based MID values can be applied across the continuum of baseline severity. However, a moderate association persists for some measures, and it is recommended that researchers continue to test for this relationship when generating anchor-based MID values from change scores. Previous suggestions that different MIDs should be applied to different strata of baseline severity or that we should use a relative measure of MID, such as percentage change, should be reconsidered.

Acknowledgments

The authors would like to acknowledge the contributions of Liz Jamieson and Lynn Copley in the data collection phase of the project. The authors also thank the Department of Health Policy Research Programme and Commercial
 Directorate for funding the project on which these analyses are based.

References

[43] Stata Corporation. Statistical software: release 9. 2007; College Station, Texas


