Responsiveness of outcome measures in patients with superior labral anterior and posterior lesions

Øystein Skare¹, Petter Mowinckel², Cecilie Piene Schrøder¹,
Sigurd Liavaag³, Olav Reikerås⁴ and Jens Ivar Brox⁵

Abstract

Background: Evaluation of patients with superior labral anterior and posterior (SLAP) lesions requires outcome measures validated for the purpose. The present study aimed to evaluate the responsiveness of the Rowe score, Oxford Instability Shoulder Score (OISS), Western Ontario Shoulder Instability Index (WOSI) and EuroQol (EQ-5D and EQ-VAS) in patients treated for SLAP lesions.

Methods: Eighty-nine patients were included: 34 had arthroscopic labral repair, 28 had mini-open biceps tenodesis and 27 had physical treatment. The outcome measures were administrated before treatment and after 6 months. Responsiveness was evaluated using standardized response mean (SRM), area under receiver-operating characteristic curve (ROCAUC), reliable chance proportion (RCP) statistics and hypothesis testing. Minimal clinically important change (MCIC) estimates were reported.

Results: All outcome measures had high values of SRM (0.86–1.92). RCPs for the improved group were 68% to 79% for OISS, WOSI and Rowe score, and 15% to 49% for EuroQol. ROCAUC was >0.70 for all outcomes. MCIC estimates were 8 and 10 for OISS; 451 and 569 for WOSI; 17 and 18 for Rowe score; 0.39 and 0.53 for EQ-5D; and 35 and 41 for EQ-VAS. Responsiveness tested with hypotheses favours the shoulder-specific outcomes.

Conclusions: OISS, WOSI and Rowe score are more responsive than EuroQol in evaluation of patients with SLAP lesions.

Keywords

Minimal clinically important change, outcome measures, responsiveness, SLAP lesions

Date received: 28th January 2014; accepted: 11th April 2014

Introduction

Isolated superior labral anterior and posterior (SLAP) lesion is a rare shoulder diagnosis compared to rotator cuff disorders.^{1–4} A SLAP lesion is a traumatic or degenerative detachment of the labrum from the glenoid, causing symptoms such as popping, clicking, pain and apprehension of glenohumeral instability. To our knowledge, there are no outcome measures developed specifically for the evaluation of patients with SLAP lesions. In a review by Gorantla et al.⁵ none of the included studies reported the use of scoring systems validated specifically for use in patients with SLAP lesions.

The usefulness of clinical scores, shoulder-specific and generic health-related quality of life outcome measures depends on their practical applicability, reliability,

¹Department of Orthopedic Surgery, Lovisenberg Diaconal Hospital, Oslo, Norway

²Department of Pediatrics, Oslo University Hospital Ullevål, Oslo, Norway

³Department of Orthopedic Surgery, Sørlandet Hospital, Arendal, Norway

⁴Department of Orthopedic Surgery, Oslo University Hospital Rikshospitalet, Oslo, Norway

⁵Department of Physical Medicine and Rehabilitation, Oslo University Hospital, Ullevål, Oslo, Norway

Corresponding author:

Øystein Skare, Lovisenberg Diaconal Hospital, Lovisenberggt. 15, 0440 Oslo, Norway.

Tel: +4792459744. Fax: +4723225514.

Email: oystein.skare@lds.no

validity and responsiveness.^{6–8} Recently published studies have reported psychometric properties in terms of agreement, reliability and validity of the 1988 version of the clinical Rowe Score, Oxford Instability Shoulder Score (OISS), Western Ontario Shoulder Instability Index (WOSI) and the EuroQol (EQ-5D and EQ-VAS) for use in patients with SLAP lesions.^{9,10} The responsiveness of these outcome measures in these patients has not been reported. Agreement was reported to be acceptable for the total scores, except for EuroQol. For the 1988 version of the Rowe score, SEM_{agreement} varied from 5.2 to 6.8, limits of agreement (LoA) were from –17.9 to 20.5 and minimal detectable change (MDC) was from 14.3 to 18.8. For OISS, SEM_{agreement} was 2.9, LoA were from –7.8 to 8.4 and MDC was 8.1. OISS had SEM_{agreement} of 122.4, LoA were from –333.9 to 344.8 and MDC was 339.3. SEM_{agreement} for EQ-5D was 0.1, LoA were from –0.06 to 0.04 and MDC was 0.4. For EQ-VAS, SEM_{agreement} was 6.0, LoA were from –17.1 to 16.2 and MDC was 16.6. There was no acceptable agreement for the domains of Rowe score and EuroQol. Reliability for the total scores as measured by interclass correlation coefficient, version 2.1 (ICC 2.1) was high for all outcome measures (ICC 2.1 > 0.70), except for EQ-5D. For the domains in EQ-5D, two domains (walking and personal care) had values <0.70. There was moderate validity for all outcome measures tested by the hypothesis method.

Responsiveness is defined as the ability of an outcome measure to detect changes over time in the construct to be measured.¹¹ There are controversies about the methodological framework for assessment of responsiveness.^{12–15} Traditionally, responsiveness has been evaluated by distribution-based methods based on statistical significance, sample variability and measurement precision. Distribution based methods include effects size (ES), standardized response mean (SRM) and MDC.¹⁶ Responsiveness depends on the treatment given and time to follow-up, and reflects change within a subject, between subjects and between groups.⁶ A statistically significant effect does not always reflect a meaningful change for the patient, clinicians or decision makers.¹⁷

By contrast, anchor-based methods estimate changes related to a meaningful external anchor.¹⁷ Such methods are used to measure the construct validity.¹⁸ The threshold for an important change is often difficult to retrieve because the statistical differences depend on sample size and measurement error.^{13,15–17} This statistical threshold is labelled minimal important change^{13,19} or minimal clinically important change (MCIC).²⁰ In the literature, the terms minimal clinically important ‘change’ and minimal clinically important ‘differences’ are often used interchangeably.²¹ A combination of distribution-based and anchor-based methods has been recommended for determining responsiveness.^{13,16,17,22}

A new strategy for the evaluation of responsiveness has been recommended by the COSMIN group. They claim that responsiveness is considered to be an aspect of the construct validity of an instrument, and should be tested by hypotheses focusing on change of scores.⁸

The purpose of the present study was to evaluate responsiveness using distribution based and anchor-based methods, and the hypothesis method of the Rowe score, the OISS, the WOSI and the EuroQol in patients treated for SLAP lesions.

Materials and methods

Study design and setting

Patients with SLAP lesions were recruited to the present study from an outpatient clinic. The inclusion criteria were a diagnosis of SLAP lesions based on clinical findings, including a positive active compression test or crank test, or a positive apprehension test with pain relief on relocation, and positive magnetic resonance image (MRI) arthrography. MRI arthrography was the main diagnostic tool²³ considering the low diagnostic value of the clinical tests.²⁴ Exclusion criteria were; patients <16 years or >60 years, former surgical repair of SLAP lesions or surgical stabilization in the same shoulder, symptomatic rotator cuff tears, symptomatic osteoarthritis in the acromioclavicular joint, osteoarthritis in the glenohumeral joint, inability to fill in the questionnaires, rheumatic disease involving the shoulder, cervical or thoracic spine pain, and major psychiatric or somatic disease. From December 2006 to August 2011, 97 patients with SLAP lesions were enrolled for inclusion in the present study. Eight were excluded from further analysis because they did not complete the anchor evaluation, leaving 89 patients for further analysis. The Rowe score was completed by one experienced physiotherapist (manual therapist) at inclusion and at 6-month follow-up. At the same time, the patients completed the WOSI, OISS, EQ-5D and EQ-VAS. Between baseline and follow-up, the patients received one of the following treatments: arthroscopic labral repair with suture anchors, mini-open bicepstenodesis with suture anchors or a standardized but individual adjusted physical rehabilitation programme. A similar rehabilitation programme with some adjustments was used in the postoperative rehabilitation. The study was approved by the Ethical Committee of Health Region South-East, Norway. All patients provided their written informed consent.

Outcome measures

Norwegian versions of the WOSI, OISS, EQ-5D and EQ-VAS have been cross-cultural adapted according to

the guidelines in the literature^{25,26} and were used in the recent study.¹⁰

The WOSI is a patient-administrated shoulder questionnaire including 21 items covering the four domains; physical symptoms (10 items); sports, recreation and work (four items); lifestyle (four items); and emotions (three items). The responses for each question are given on a scale from 0 to 100, with 0 as the best possible score and 2100 as the worst. The worst possible total score 2100 indicates an extremely low health-related quality of life related to the shoulder.²⁷

The OISS is a patient administrated shoulder questionnaire including 12 questions with four response alternatives (1 point to 5 points). The total score ranges from 12 to 60, with 12 indicating best possible function.²⁸

The EuroQol (EQ-5D and EQ-VAS) is a generic patient administrated health-related quality of life instrument. EQ-5D covers five domains: mobility; self-care; usual activities; pain/discomfort; and anxiety/depression. There are three response alternatives: no problem; some problem; and extreme problem. Responses are transformed into an index, where 1.0 represents the highest level of function and -0.59 represents the worst possible state. EQ-VAS is a visual analogue scale measuring the current health state, ranging from 0 (worst imaginable health status) to 100 (best imaginable health status).²⁹⁻³¹

The 1988 version of the Rowe Score³² is the fourth version of the Rowe score. The original version of the Rowe score³³ was developed for the evaluation of patients with Bankart lesions. The 1988 version consists of five domains: pain; stability; function; motion; and muscle strength. Pain and stability scores are weighted differently depending on which condition is considered to be the major problem. The best possible total score is 100, representing a normal shoulder. Total scores may be categorized into four levels: excellent (100 points to 85 points); good (84 points to 70 points); fair (69 points to 50 points); and poor (≤ 49 points).

Anchors for important change

In addition, the patients assessed their shoulder at 6-month follow-up as excellent, good, fair or poor in the patient evaluation of 1988 version of the Rowe score.³² Excellent or good were merged and labelled improved, whereas fair and poor were labelled unimproved. They also scored their main change of symptoms at 6-month follow-up on an 18-point scale ranging from -9 (worst possible) to 9 (best possible).³⁴

Statistical analysis

Sample size. The sample size of at least 50 subjects was chosen according to recommendations for estimation of MCIC³⁵ and for method comparison studies.³⁶

Descriptive analysis. Nonparametric Spearman rho correlations were obtained between the anchor of important improvement and change in main complaint (-9 to 9).

Responsiveness. Responsiveness was assessed using SRM; receiver-operating curve (ROC) analysis and reliable change proportion (RCP).

SRM was calculated for the improved group (SRM_{improved}) and the unimproved group ($SRM_{\text{unimproved}}$), dividing the mean change scores by the SD of mean change scores.³⁷ SRM values of 0.20, 0.50 and 0.80 are obtained in the same way as for ES, and are considered to be small, moderate or large, respectively.³⁸⁻⁴⁰ A 95% confidence interval (CI) for SRM was estimated using a bootstrap procedure.

ROC curve analysis was used at 6-month follow-up to evaluate the accuracy of each instrument (i.e. the sensitivity and specificity of a true classification as improved or unimproved linked to a the anchor of main symptoms) (18-point scale described above). A ROC curve is a plot of sensitivity against 1 - specificity of the false positive rate. Each point on the scale (operating points) defines different cut-off levels for trade-off between true-positive success and false negative error. The area under the ROC curve (ROC_{AUC}) describes the ability of the instrument to distinguish between patients who are improved or unimproved. A value of 1.00 indicates a perfect ability to discriminate improved from unimproved patients, whereas a value of 0.50 indicates that the ability is no better than chance. The ROC_{AUC} is reported with 95% CI.

RCP is defined as the proportion (%) of patients improved by more than the smallest detectable change/MDC.⁴¹ The MDC for an individual patient with 95% CI was calculated for each outcome using the formula $MDC = SEM_{\text{agreement}} \times 1.96 \times \sqrt{2}$, where $SEM_{\text{agreement}}$ is the standard error of measurement and 1.96 is the 95% level of confidence. The value for $SEM_{\text{agreement}}$ was obtained by extracting the square root of the residual mean square for change, using one-way analysis of variance with subjects as the factor.^{42,43} The estimates of MDC were obtained from 71 patients completing the outcome measures twice at the same time of day with a 1-week interval.^{9,10}

MCIC. There are several different strategies to define minimal clinically important change.^{13,15,17,35} There is little evidence for an explicit MCIC value in relation to the method applied.⁴⁴ In the present study, minimal

clinically important change was estimated using the anchor-based minimal clinically important change distribution method described by de Vet et al.¹³ MCIC was assessed for the improved group using different cut-off points: the optimal cut-off value (95% CI bootstrap value), the ROC cut-off point and the 95% limit cut-off point. The MCIC is not a single or a fixed value and is reported in the present study as a set of values as recommended by de Vet et al.¹³

The optimal cut-off point is a measure of the accuracy to sample estimates; in this case, to discriminate between improved and unimproved according to an anchor. The ROC-cut-off point was retrieved from ROC analysis. The MCIC using the ROC analysis (MCIC_{ROC}) was determined as the upper left point in the ROC curve for each outcome instrument. The 95% limit cut-off point (MCIC_{95%limit}) was calculated from the distribution of the patients who are not changed significantly, according to the anchor. The rationale is that MCIC should be detected beyond measurement error. MCIC_{95%limit} is defined as the 95% upper limit of the distribution of patients who are not changed significantly (mean change + 1.645 SD_{change}). The MCIC_{95%limit} equals the 95% specificity on the ROC curve.¹³

Hypotheses. Responsiveness should be evaluated similar to criterion validity because the only difference between criterion validity and responsiveness is that validity refers to the validity of a single score, and responsiveness refers to the validity of change of scores. Hypothesis testing should focus on the change scores of an instrument and should be predefined, including the expected direction of correlations or differences in change.⁸ In the present study, a correlation >0.70 was regarded as positive.³⁵ The responsiveness was tested using the hypotheses:

1. Mean change scores for women and men should be statistically indifferent.
2. Mean change scores of patients <40 years and >40 years should be statistically indifferent.
3. The mean change scores and the change in the main complaint (−9 to 9) should be correlated >0.70.
4. The mean change scores and the categorical change scores (improved or unimproved) should be correlated >0.70.
5. The 6-month follow-up scores should be correlated <0.70 with the categorical patient outcome (improved or unimproved).
6. The 6-months follow-up scores should be correlated <0.70 with the change in main complaint (−9 to 9).
7. The mean change scores of patients with and without postoperative stiffness should be statistically different.
8. The mean change scores of the outcome measures should be correlated >0.70.

Hypotheses (1), (2) and (7) were tested using an independent sample *t*-test. For hypotheses (3) to (6), we used the Spearman rank correlation because the change in main complaint and the categorical change score was not normally distributed. Hypothesis (8) was tested using Pearson *r* correlation because the scores were normally distributed.

The analyses were performed using SAS, version 9.3 (SAS Institute Inc., Cary, NC, USA) and SPSS, version 18 (SPSS, Chicago, IL, USA). The exact binomial distribution was used to calculate the confidence interval of the binomial estimates.

Results

Fifty-seven (64%) males and 32 (36%) females with a median age of 39 years (range 18 years to 60 years) were included for further analysis in the present study (Table 1). Two patients did not complete all items in the WOSI at 6 months, although they were included for analysis of the other outcome measures. Fifty-nine patients (66%) were labelled improved and 30 (34%) were labelled unimproved according to the anchor.

Table 1. Descriptive statistics.

Males/females	57/32
Age (years), mean (range)	39.4 (18 to 60)
Duration of symptoms (months), median (range)	27 (4 to 192)
Manual labour	
Physical activity	45 (51)
Competition	7 (8)
Weekly or more	44 (49)
None	38 (43)
Dominant shoulder involved	67 (75)
Postoperative capsular stiffness	17 (19)
Treatment	
Arthroscopic labral repair	34 (38)
Mini-open biceps tenodesis	28 (32)
Rehabilitation	27 (30)

The number (%) of patients is given if not reported otherwise.

All change of scores had a normal distribution. There were no floor and ceiling effects of the total scores of the outcome measures at baseline. The correlation between the categorical change scores (improved or unimproved) and the change in main complaint (−9 to 9) and the mean change scores was 0.84.

Distribution-based responsiveness

SRM with 95% CI and mean change scores for improved and unimproved patients are reported in Table 2. The SRM values were high for all six scores in the improved group and low for all outcome measures in the unimproved group. RCPs for the improved group were 68% and 73% for the Rowe score; 79% for WOSI; 73% for OISS; 15% for EQ-5D; and 49% for EQ-VAS (Table 2). In the unimproved group, the RCPs ranged from 17% (EQ-5D) to 45% (WOSI).

Anchor-based responsiveness

Fifty-nine patients (66%) were categorized as improved according to the anchor. The results of the ROC analysis are reported in Fig. 1 and Table 3. The ROC_{AUC} was higher than 0.70 for all outcome measures. The distribution of the mean change scores according to the anchor is reported in Fig. 2.

Minimal clinically important change

The cut-off values for minimal clinically important change are given in Table 3. MCIC_{ROC} and MCIC_{95%limit} was 10 and 8 for OISS; 569 and 451 and for WOSI; 17 for Rowe score; 0.53 and 0.59 for EQ-5D; and 35 and 41 for EQ-VAS. When transforming of the scores into 0 to 100 to ease the comparison between score, the MCIC_{ROC} and MCIC_{95%limit} was 16.7 and 13.4 for OISS; 27.1 and 21.5 for WOSI; 17 for Rowe score; 53 and 59 for EQ-5D; and 35 and 41 for EQ-VAS. The ROC cut-off value using a bootstrap procedure was substantially lower than the values for MCIC_{ROC} and MCIC_{95%limit}.

Hypotheses

There were no statistically significant differences in any change scores between men and women and between patients aged <40 years or aged >40 years (Hypothesis 1 and 2). The correlations between the mean change scores and change in main complaint (Hypothesis 3) were 0.78 for the OISS and WOSI; 0.71 and 0.72 for Rowe score; 0.55 for EQ-5D; and 0.57 for EQ-VAS. Mean change of scores and the categorical change scores (improved or unimproved) correlated 0.67 for OISS; 0.65 for WOSI; 0.69 and 0.70 for Rowe score;

0.49 for EQ-5D; and 0.45 for EQ-VAS (Hypothesis 4). The correlations between the outcome scores at 6-month follow-up and the categorical change score (improved or unimproved) were 0.59 for OISS; 0.60 for WOSI; 0.66 and 0.68 for Rowe score; 0.54 for EQ-5D; and 0.37 for EQ-VAS (Hypothesis 5). Change in main complaint and the outcome scores at 6-month follow-up correlated 0.66 for OISS; 0.75 for WOSI; 0.70 and 0.72 for Rowe score; 0.63 for EQ-5D; and 0.58 for EQ-VAS (Hypothesis 6).

The mean change scores for patients with and without postoperative capsular stiffness were statistically different only for the Rowe score (Hypothesis 7). The outcome measures mean change scores (Hypothesis 8) for OISS correlated 0.84 with WOSI; 0.66 with Rowe score; and 0.40 with EQ-5D and 0.53 with EQ-VAS. WOSI correlated 0.71 and 0.58 with Rowe score and EuroQol, respectively. The correlations for Rowe score with EQ-5D and EQ-VAS were 0.45 and 0.48, respectively.

Discussion

The present study aimed to evaluate the responsiveness of scores used for the evaluation of outcome in patients with SLAP lesions.⁵ Using strategies comprising distribution-based methods, anchor-based methods and the hypothesis method, we found that the self-reported OISS, WOSI and the clinician-assessed Rowe score are responsive in the evaluation of patients with SLAP lesions. The OISS and Rowe score appear to be the most sensitive measures. The anchor-based results suggest that the generic EuroQol is less responsive. The results suggest that the anchors discriminated well between those who considered themselves improved compared to those who did not. The distribution-based methods using SRM and RCP indicate relatively small differences in responsiveness between the clinical-assessed Rowe score and self-reported shoulder-specific questionnaires OISS and WOSI. By contrast, the anchor-based methods, MCIC and ROC analysis, and hypotheses testing indicate that the Rowe score and the shoulder-specific questionnaires are more responsive than the generic EuroQol.

Distribution-based responsiveness measured with SRM and RCP were in favour of the shoulder-specific outcome measures. The WOSI had the highest value of RCP, whereas the Rowe score was superior in SRM. The area under the ROC curve where sensitivity and specificity is weighted equally importantly showed that all questionnaires were responsive. The best results were obtained for the shoulder-specific questionnaires, OISS and WOSI.

The MCIC estimates indicates that change of scores needs to be >17 for Rowe score; >569 for WOSI; >10

Table 2. Mean baseline scores, change scores, SD, standardized response mean (SRM) and reliable proportion change (RCP) with 95% confidence interval (CI) in patients rating their shoulder as improved (excellent or good,) and unimproved (fair or poor) at 6 months.

Improved			Unimproved									
	Rowe 25	Rowe 15	WOSI	OISS	EQ-5D	EQ- VAS	Rowe 25	Rowe 15	WOSI	OISS	EQ-5D	EQ VAS
Baseline score	61.8	66.0	1122.6	38.3	0.65	67.5	65.4	69.2	970.0	35.6	0.67	70.6
95% CI	(59.5 to 64.2)	(63.8 to 68.2)	(1035.6 to 1209.6)	(36.5 to 40.0)	(0.6 to 0.7)	(63.7 to 71.4)	(61.6 to 69.1)	(65.5 to 73.0)	(760.7 to 1189.3)	(32.4 to 38.9)	(0.6 to 0.8)	(61.1 to 79.4)
SD	10.1	9.5	373.0	7.5	0.22	16.4	7.1	7.0	394.3	6.0	0.22	16.7
Change score	−20.0	−17.3	654.0	11.8	−0.17	−14.8	1.9	2.7	45.4	−0.2	0.04	5.6
95% CI	(−23.9 to −16.0)	(−20.8 to −13.8)	(540.2 to 767.8)	(9.7 to 14.0)	(−0.22 to 0.11)	(−18.7 to −10.8)	(−2.86 to 6.73)	(−2.06 to 7.43)	(−84.4 to 175.3)	(−6.7 to 23)	(−0.07 to 0.15)	(−6.0 to 17.1)
SD	16.9	14.8	480.9	9.1	0.25	16.9	9.0	9.0	244.8	4.7	0.21	21.7
SRM	1.92	1.80	1.73	1.65	0.86	1.06	−0.12	−0.13	0.35	0.23	−0.05	−0.08
95% CI	(1.49 to 2.35)	(1.38 to 2.21)	(2.14 to 1.32)	(1.26 to 2.04)	(0.56 to 1.16)	(0.74 to 1.38)	(−0.48 to 0.24)	(−0.49 to 0.23)	(−0.03 to 0.72)	(−0.14 to 0.59)	(−0.41 to 0.31)	(−0.44 to 0.28)
RCP	73%	68%	79%	73%	15%	49%	23%	30%	45%	43%	17%	33%
95% CI	(60 to 84%)	(54 to 79%)	(67 to 89%)	(60 to 84%)	(7 to 27%)	(36 to 63%)	(10 to 42%)	(15 to 49%)	(26 to 64%)	(25 to 63%)	(6 to 35%)	(17 to 53%)

Rowe 25, Rowe Score 1988 version using 25 pain points; Rowe 15, using 15 pain points; WOSI, Western Ontario Shoulder Instability Index; OISS, Oxford Instability Shoulder Score; EQ-5D, EuroQol 5 Domains; EQ VAS, EuroQol VAS scale.

for OISS; >0.59 for EQ-5D; and >41 for EQ-VAS to be considered clinically important using the most strict estimates. For the Rowe score, the estimate is approximately the same as the measurement error measured as the MDC/repeatability coefficient previously reported in a test-retest study.⁹ The measurement error was reported to be considerably lower than the MCIC estimates for OISS, WOSI, EQ-5D and EQ-VAS.¹⁰ The MCIC estimates suggests that OISS and Rowe score

are the most sensitive outcome measures because a lower score in these outcomes is needed to detect clinically important changes. The superior results for OISS and Rowe score are also demonstrated by the sensitivity measures in Table 3. The distribution-based method may overestimate the responsiveness of EQ-5D, whereas the anchor-based methods suggests that EQ-VAS is not responsive. A possible explanation is that generic health-related quality of life outcomes are more influenced by general health and co-morbidities than shoulder-specific outcomes.

As noted by de Vet et al.¹³, the value for improvement and deterioration is not always the same, and they suggest defining MCIC as the change that patients would consider important for reaching their situation, depending on severity of disease, type of intervention and duration of follow-up period. The MCIC value is considered important for clinicians both as a benchmark for the evaluation of clinical significance and for use in power calculations.^{13,15,19,29} In randomized controlled trials, estimates of clinical importance are essential for power calculations. Even small differences can be recognized if the measure outcome is sensitive in clinical trials with large samples. These changes may be clinically irrelevant. By contrast, if the outcome measure has low sensitivity or the sample is small, clinically important differences may be missed.

One advantage of the anchor-based method is the possibility to detect change from the patient perspective. Another is the ability to use other relevant information, such as baseline scores and regression to the mean, which provide information about the tendency

Figure 1. Receiver-operating curve (ROC) analysis. diff, differences between test and re-test; row25/15, Rowe score 1988 version with the weighting of 25 or 15 pain points; eq5dscore: total cores of EuroQol-5 domains; eq5dvas, EuroQol-VAS; wosi, Western Ontario Shoulder Instability Index; oiss, Oxford Instability Shoulder Score.

Table 3. Area under receiver-operating curve (ROC) and cut off-value with 95% confidence interval (CI), sensitivity and specificity, minimal detectable change (MDC) and minimal clinically important change (MCIC) estimates for patients rating important change in main complaint.

	ROC _{AUC} (95% CI)	ROC cut-off value (95% CI)	Sensitivity/specificity	MDC	MCIC _{ROC}	MCIC _{95%limit}
OISS	0.92 (0.84 to 0.96)	5 (4 to 7)	0.88/0.83	8.1	10	8
WOSI	0.90 (0.81 to 0.94)	205 (136 to 458)	0.79/0.93	339.3	569	451
Rowe 25	0.93 (0.85 to 0.97)	3 (−1 to 14)	0.93/0.83	14.3	17	17
Rowe 15	0.93 (0.84 to 0.97)	5 (−2 to 13)	0.91/0.86	16.0	17	18
EQ-5D	0.81 (0.68 to 0.88)	0.035 (0.000 to 0.069)	0.74/0.86	0.4	0.53	0.39
EQ-VAS	0.79 (0.66 to 0.87)	4 (−5 to 10)	0.72/0.76	16.6	35	41

Numbers are given as raw scores in the scale of the different instruments. ROC_{AUC}, area under receiver-operating characteristic curve; cut off-value designation of a positive or negative outcome from bootstrap analyses; MDC, minimal detectable change from test-retest study; MCIC_{ROC}, minimal clinically important change from cut-off point in receiver operator characteristic curve; MCIC_{95%limit}, minimal clinically important change from 95% limit of the unimproved (fair/poor) group. Rowe 25, Rowe Score 1988 version using 25 pain points; Rowe 15, using 15 pain points; WOSI, Western Ontario Shoulder Instability Index; OISS, Oxford Instability Shoulder Score; EQ-5D, EuroQol 5 Domains; EQ VAS, EuroQol VAS scale.

Figure 2. Boxplots mean change scores improved versus unimproved.

of extreme scores to become more normal at follow-up.⁴⁵ A disadvantage of the anchor-based methods is recall bias,⁴⁶ which implies that the change scores reflect the condition at follow-up, and not just the change from baseline. In addition, there is little knowledge about the reliability and validity of anchor-based methods. It is also a disadvantage that the anchor-based methods do not reflect the measurement precision (i.e. a clinical meaningful change may be within the random variation of the instrument).^{16,45} For these reasons, a combination of distribution-based methods and anchor-based methods to establish responsiveness is preferred.^{13,15,45}

Hypothesis testing suggests that responsiveness is independent of gender and age. The good correlations between the mean change of scores and the change in main complaint for OISS, WOSI and Rowe score indicate that these outcomes are more responsive than EuroQol. None of the outcome measures correlated well with the categorical outcome, which may be explained by the fact that this anchor is categorical. The correlation between the mean change of scores and the change scores at follow-up gives the opportunity to check for recall bias. The results do not exclude recall bias for OISS, WOSI and Rowe score. The Rowe score was superior to the other outcomes in differentiating between patients with or without capsular stiffness, which is possibly attributed to measuring the range of movement. The correlations were generally better for the shoulder-specific OISS and WOSI than for the generic EQ-5D and EQ-VAS (EuroQol), suggesting that the generic outcomes are measuring a different construct. The major concerns of the use of hypotheses for the evaluation of responsiveness are that the choice of hypotheses may be fitted for the wanted conclusion and the results, especially if they are not reported in a protocol before publishing the results.³⁵

The main strength of the present study is the use of different strategies, allowing the clinician to obtain an overview of both the distribution-based and anchor-based responsiveness for the current population. Rather than a fixed value, the MCIC is presented as a range of values obtained from different methods. This will provide the opportunity to make a judgement of the threshold for important change depending on the chosen anchor. Hypotheses were used as recommended.^{8,11} In addition, patients receiving operative and non-operative treatment were included, which suggests that the results are valid for these treatments.

One of the limitations of the present study is the size of the study population. Although the power of the study is sufficient according to recommendations of Altman et al.,³⁶ and Terwee et al.³⁵ the study population was considered to be too small to allow for subgroup analysis of deteriorated patients as

recommended.¹³ This would have been valuable because the MCIC values for deterioration and improvement are likely to be different, and deterioration is a possible outcome after treatment. Another limitation is that the responsiveness and MCIC is measured only at 6-month follow-up. The use of a shorter or longer follow-up would provide valuable information about the influence of time on responsiveness.

Conclusions

The self-reported OISS, WOSI and the clinician-assessed Rowe score are responsive in the evaluation of patients with SLAP lesions. The OISS and Rowe score appear to be the most sensitive measures. The anchor-based results suggest that the generic questionnaires are less responsive. The advantage of the self-reported questionnaires is that they are easier to administer.

Declaration of conflicting interests

None declared.

Funding

This study was supported by Lovisenberg Diaconal Hospital, Oslo, Norway.

References

1. Kim TK, Queale WS, Cosgarea AJ and McFarland EG. Clinical features of the different types of SLAP lesions: an analysis of one hundred and thirty-nine cases. *J Bone Joint Surg Am* 2003; 85A: 66–71.
2. Snyder SJ, Karzel RP, Del PW, Ferkel RD and Friedman MJ. SLAP lesions of the shoulder. *Arthroscopy* 1990; 6: 274–9.
3. Snyder SJ, Banas MP and Karzel RP. An analysis of 140 injuries to the superior glenoid labrum. *J Shoulder Elbow Surg* 1995; 4: 243–8.
4. Weber SC, Martin DF, Seiler JG III and Harrast JJ. Superior labrum anterior and posterior lesions of the shoulder: incidence rates, complications, and outcomes as reported by American Board of Orthopedic Surgery. Part II candidates. *Am J Sports Med* 2012; 40: 1538–43.
5. Gorantla K, Gill C and Wright RW. The outcome of type II SLAP repair: a systematic review. *Arthroscopy* 2010; 26: 537–45.
6. Beaton DE, Bombardier C, Katz JN and Wright JG. A taxonomy for responsiveness. *J Clin Epidemiol* 2001; 54: 1204–17.
7. Kirshner B and Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis* 1985; 38: 27–36.
8. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol* 2010; 10: 22.

9. Skare O, Schroder CP, Mowinckel P, Reikeras O and Brox JI. Reliability, agreement and validity of the 1988 version of the Rowe Score. *J Shoulder Elbow Surg* 2011; 20: 1041–9.
10. Skare O, Liavaag S, Reikeras O, Mowinckel P and Brox JI. Evaluation of Oxford instability shoulder score, Western Ontario shoulder instability index and Euroqol in patients with SLAP (superior labral anterior posterior) lesions or recurrent anterior dislocations of the shoulder. *BMC Res Notes* 2013; 6: 273.
11. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010; 63: 737–45.
12. Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness. *BMC Med Res Methodol* 2011; 11: 152.
13. de Vet HC, Ostelo RW, Terwee CB, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res* 2007; 16: 131–42.
14. Terwee C, Dekker F and Bossuyt P. A taxonomy for responsiveness? *J Clin Epidemiol* 2002; 55: 1156.
15. Terwee CB, Roorda LD, Knol DL, de Boer MR and de Vet HC. Linking measurement error to minimal important change of patient-reported outcomes. *J Clin Epidemiol* 2009; 62: 1062–7.
16. Crosby RD, Kolotkin RL and Williams GR. An integrated method to determine meaningful changes in health-related quality of life. *J Clin Epidemiol* 2004; 57: 1153–60.
17. Crosby RD, Kolotkin RL and Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003; 56: 395–407.
18. Lydick E and Epstein RS. Interpretation of quality of life changes. *Qual Life Res* 1993; 2: 221–6.
19. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL and Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* 2006; 4: 54.
20. de Vet HC, Beckerman H, Terwee CB, Terluin B and Bouter LM. Definition of clinical differences. *J Rheumatol* 2006; 33: 434.
21. Jaeschke R, Singer J and Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989; 10: 407–15.
22. Revicki D, Hays RD, Cella D and Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008; 61: 102–9.
23. Chang D, Mohana-Borges A, Borso M and Chung CB. SLAP lesions: anatomy, clinical presentation, MR imaging diagnosis and characterization. *Eur J Radiol* 2008; 68: 72–87.
24. Hegedus EJ, Goode AP, Cook CE, et al. Which physical examination tests provide clinicians with the most value when examining the shoulder? Update of a systematic review with meta-analysis of individual tests. *Br J Sports Med* 2012; 46: 964–78.
25. Beaton DE, Bombardier C, Guillemin F and Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila Pa 1976)* 2000; 25: 3186–91.
26. Guillemin F, Bombardier C and Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol* 1993; 46: 1417–32.
27. Kirkley A, Griffin S, McLintock H and Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). *Am J Sports Med* 1998; 26: 764–72.
28. Dawson J, Fitzpatrick R and Carr A. The assessment of shoulder instability. The development and validation of a questionnaire. *J Bone Joint Surg Br* 1999; 81: 420–6.
29. Brazier J, Jones N and Kind P. Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. *Qual Life Res* 1993; 2: 169–80.
30. Brooks R. EuroQol: the current state of play. *Health Policy* 1996; 37: 53–72.
31. Essink-Bot ML, Krabbe PF, Bonsel GJ and Aaronson NK. An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. *Med Care* 1997; 35: 522–37.
32. Rowe CR. Evaluation of the shoulder. In: Rowe CR (ed.) *The shoulder*. New York, NY: Churchill Livingstone, 1988, pp.631–7.
33. Rowe CR, Patel D and Southmayd WW. The Bankart procedure: a long-term end-result study. *J Bone Joint Surg Am* 1978; 60: 1–16.
34. Ekeberg OM, Bautz-Holter E, Keller A, Tveita EK, Juel NG and Brox JI. A questionnaire found disease-specific WORC index is not more responsive than SPADI and OSS in rotator cuff disease. *J Clin Epidemiol* 2010; 63: 575–84.
35. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007; 60: 34–42.
36. Altman DG. *Practical statistics for medical research*. London: Chapman & Hall, 1991:402–3.
37. Guyatt GH, Bombardier C and Tugwell PX. Measuring disease-specific quality of life in clinical trials. *CMAJ* 1986; 134: 889–95.
38. Beaton DE, Hogg-Johnson S and Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. *J Clin Epidemiol* 1997; 50: 79–93.
39. Garratt AM, Ruta DA, Abdalla MI and Russell IT. SF 36 health survey questionnaire: II. Responsiveness to changes in health status in four common clinical conditions. *Qual Health Care* 1994; 3: 186–92.
40. Liang MH, Fossel AH and Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990; 28: 632–42.

41. Schmitt JS and Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol* 2004; 57: 1008–18.
42. Bland JM and Altman DG. Measurement error. *BMJ* 1996; 313: 744.
43. Hinton-Bayre AD. Specificity of reliable change models and review of the within-subjects standard deviation as an error term. *Arch Clin Neuropsychol* 2011; 26: 67–75.
44. Beaton DE, Boers M and Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol* 2002; 14: 109–14.
45. Crosby RD, Kolotkin RL and Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003; 56: 395–407.
46. Grovle L, Haugen AJ, Keller A, Natvig B, Brox JI and Grotle M. Poor agreement found between self-report and a public registry on duration of sickness absence. *J Clin Epidemiol* 2012; 65: 212–8.