Skare *et al. BMC Musculoskeletal Disorders* (2021) 22:606
https://doi.org/10.1186/s12891-021-04483-3

BMC Musculoskeletal Disorders

RESEARCH Open Access

Responsiveness of five shoulder outcome measures at follow-ups from 3 to 24 months

Øystein Skare[1,2*], Jostein Skranes Brox[3], Cecilie Piene Schrøder[1] and Jens Ivar Brox[4,5]

Abstract

Background: To assess responsiveness of five outcome measures at four different follow-ups in patients with SLAP II lesions of the shoulder.

Methods: 119 patients with symptoms and signs, MRI arthrography and arthroscopic findings were included. The Western Ontario Shoulder Instability Index (WOSI), Oxford Instability Shoulder Score (OISS), EuroQol (EQ-5D3L), Rowe Score and Constant-Murley Score (CMS) were assessed at baseline, 3, 6, 12 and 24 months. The analysis contains both anchor-based and distribution-based methods, and hypothesis testing.

Results: Confidence intervals for ROC cut-off values, representing MID, for OISS, CMS and EQ-5D3L crossed zero at 3 months. Cut-off values were stable between 6- and 24-months follow-up. At 24-months ROC cut-off values (95% CI) were: Rowe 18 (13 to 24); WOSI 331 (289 to 442); OISS 9 (5 to 14); CMS 11 (9 to 15) and EQ-5D3L 0.123 (0.035 to 0.222). $\mathrm{MID}_{95\%\mathrm{limit}}$ estimates were substantially higher than ROC cut-off values and $\mathrm{MID}_{\mathrm{MEAN}}$ at all follow-ups for all instruments. The reliable change proportion (RCP) values in the improved group were highest for WOSI and the Rowe Score (ranging from 68 to 87%) and significantly lower for CMS. EQ-5D3L had the lowest values (13 to 16%). We found a moderate correlation between mean change scores of the outcome measures and the anchor, except for the EQ-5D3L.

Conclusions: In patients with SLAP II-lesions the patient reported OISS and WOSI and the clinical Rowe score had best responsiveness. Our results suggest that 3 months follow-up is too early for outcome evaluation.

Keywords: Minimal important difference, MCID, Responsiveness, SLAP, Shoulder, EQ-5D3L, Clinical scores, PROMS

Background

Clinical scores, and patient reported outcome measures (PROMS) are recommended for evaluation of treatment effects in patients with shoulder disorders. Over the last decades, the patient perspective has been considered increasingly important [1] and shoulder specific and generic health related quality of life outcome measures may replace clinical scores [2]. The shoulder specific questionnaires commonly include questions about pain and activities in daily life. The questionnaires are developed for shoulder patients in general or for specific subgroups, for example patients with instability. The answers are transformed into metrics and the scientific field handling the corresponding issues are labelled psychometrics or clinimetrics [3, 4]. Researchers and clinicians should keep in mind that some information is lost when pain and disability are transformed into metrics and that responsiveness is a word used to describe sports cars [5]. Clinimetric responsiveness describes the ability

* Correspondence: oystein.skare@lds.no
[1]Surgical Department, Lovisenberg Diaconal Hospital, Oslo, Norway
[2]Orthopedic Department, Lovisenberg Diaconal Hospital, Lovisenberggt 17, 0440 Oslo, Norway
Full list of author information is available at the end of the article

of an outcome measure to detect a change that is not at random.

Generic quality of life questionnaires, like the EuroQol (EQ-5D3L), are often applied as an utility index in cost effectiveness studies [6], despite the fact that its reliability and usefulness in shoulder patients have been questioned [7]. While a possible advantage of questionnaires is that they may be answered online to save time and costs of consultations and travel, the clinical scores provide additional information about range of motion, muscle strength and stability that may be important for the clinician and the patient. However, observations are prone to blinding and inter- and intra-rater measurement error.

An outcome measure in patients with a shoulder disorder should be evaluated for reliability and validity. Responsiveness refers to the validity of change scores while criterion validity refers to the validity of a single score. Hypothesis testing is recommended to assess validity. Hypotheses should be predefined and assess the direction and difference in change. The minimal important difference (MID) provides the clinician with an estimate of the difference between no change and minimal change on the outcome measures according to patient perceived improvement and is commonly used as a measure of responsiveness. The MID is different from a moderate or large treatment effect and the proportion of patients above MID in a study is not equivalent with the success rate. There are several methodological concerns, for example: What is the best follow-up time for a valid assessment of MID? Are the measures of responsiveness different for clinical scores and PROMS? How should MID be estimated, and how should uncertainty both in measurement and in methodology be considered?

There are no outcome measures developed specifically for patients treated for SLAP- lesions [8]. The 1998 version of the Rowe Score, WOSI, OISS and EQ-5D3L have previously been validated at 6 months [7–9]. The CMS is a recommended clinical score for all types of shoulder disorders but has not been validated for use in patients with SLAP lesions [10]. The aim of the present study was to compare the responsiveness of these five different outcome measures at four different follow-ups (3, 6, 12 and 24 months) in patients with SLAP II lesions using both anchor- and distribution-based methods. The second aim was to evaluate validity of the outcome measures by hypothesis testing.

Methods

Study design and settings

This is a prospective methodology study combining the use of distribution- and anchor-based methods, and hypotheses to assess responsiveness of outcome measures in patients with type II SLAP lesions from 3 to 24 months follow-up.

Patients were recruited from the outpatient clinic at the Department of Orthopaedic Surgery at Lovisenberg Diaconal Hospital between January 2008 and January 2014. They were originally enrolled in a blinded, three-armed, randomized, sham-controlled study with a 24 months follow-up assessing the clinical effectiveness of arthroscopic labral repair and biceps-tendinosis in patients with type II SLAP lesions [11, 12]. Some of these patients were included in a study of responsiveness at 6 months follow-up [9]. The patients enrolled were between 18 and 60 years old and had experienced shoulder pain and disability for at least 3 months prior to inclusion, despite having received non-operative treatment (physical therapy, non-steroid anti-inflammatory drugs, corticosteroid injections). They had symptoms, clinical findings and MR –arthrography indicating type II SLAP lesion. The diagnosis was confirmed during arthroscopy. The inclusion- and exclusion criteria, and a flow chart, have been described in detail previously [11, 13]. We obtained written informed consent from all participants. Ethics approval (IRB00001870) was received from the Ethics Committee Health Region Southeast, Oslo, Norway. The protocol was registered at ClinicalTrials.gov (NCT00586742).

Patients enrolled in the study completed the WOSI, OISS and EQ-5D3L at baseline and at 3, 6, 12- and 24-months follow-up. Comparisons between groups have been published previously [11]. The clinical Rowe score (1988-version) and CMS were completed by one single experienced Manual Therapist (ØS) at all follow-ups.

Outcome measures

We used previously translated and validated Norwegian versions of the patient reported WOSI [14], OISS [15] and EQ-5D3L [7, 16]. All outcome measures including the Rowe score and CMS are described in Table 1 [17, 18].

Anchor for important change

The anchor was an assessment of change of symptoms on a continuous scale ranging from - 9 (worst possible change of symptoms) to 9 (best possible) and Rowe patient evaluation [19]. This is a question were patients state their shoulder as “Excellent”, “Good”, “Fair” or “Poor”. To divide patients into groups of improved/unchanged/deteriorated both questions were combined using distribution plots. This is further described in the statistics section and in Fig. 1.

Sample size

Sample size was calculated to accommodate the RCT and was above the general recommendations for estimation of MID [11, 20, 21].

Table 1 Outcome measures

Outcome measure	Domains	Score
Patient rated outcome measures		
WOSI*	21 questions answered on a continuous scale from 0 to 100 covering physical symptoms; sports; recreation and work; lifestyle; and emotions	0 to 2100 (worst possible condition)
OISS*	12 questions with five response alternatives (5 to 0 points) covering instability, daily activities, pain, work, social life, sports/hobbies, attention to the shoulder problem, lifting and lying position	12 to 60 (worst possible function)
EQ-5D	5 questions with three response categories covering mobility; self-care; usual activities; pain/discomfort; and anxiety/depression	−0.59 to 1.0 (best quality of life)
Clinician evaluated outcome measures		
Rowe score#	5 domains evaluated: pain; stability; function, range of motion measured by a goniometer; and muscle strength using a spring gauge. Pain (25 points) and function were reported on five-step categorical scales.	0 to 100 (best possible state)
Constant-Murley Score	4 domains evaluated: pain intensity; activities of daily living (sleep, work, recreation and hand positioning); active range of motion measured by goniometer and strength of abduction measured with a spring gauge.	0 to 100 (best possible shoulder function)

* For statistical analyses, WOSI and OISS scores were inverted for easier comparisons with the other outcome measures

# At the follow-up examinations, patients were told to rate the state of their shoulder as excellent, good, fair or poor (Rowe Patient Evaluation)

Statistics

Total scores for all outcome measures were calculated at all follow-ups. For WOSI missing values were imputed if one or two questions were missing, using the mean value within each subcategory for the given patient. For CMS and OISS we imputed the mean value of the given question. All together 13 observations were imputed.

Responsiveness

Responsiveness was calculated and investigated using SRM (Standardized Response Mean), RCP (Reliable Change Proportion) and Receiver operating characteristic (ROC) at all follow-ups. The improved and unchanged group were defined by the anchor described above. Cut-off values were decided using distribution plots of the change of symptoms anchor grouped by a dichotomized version of Rowe Patient Evaluation (Fig. 1). This yielded cut-off values on the change of symptoms anchor of 6, 4, 3 and 3 at the different follow-ups. Patients scoring below − 3 were considered deteriorated and excluded from the analysis [22]. Patients with a score between − 3 and cut-off were considered unchanged. We chose not to estimate MID for the deteriorated group due to the small sample size (ranging from 2 to 16).

SRM was estimated by dividing the MCS (Mean Change Score) by the standard deviation of the MCS. 95% confidence intervals were obtained by non-parametric bootstrap estimation. Confidence intervals for baseline scores and MCS were estimated using the normal distribution. All confidence intervals for EQ-5D3L were obtained by non-parametric bootstrap estimation due to non-normal distributions.

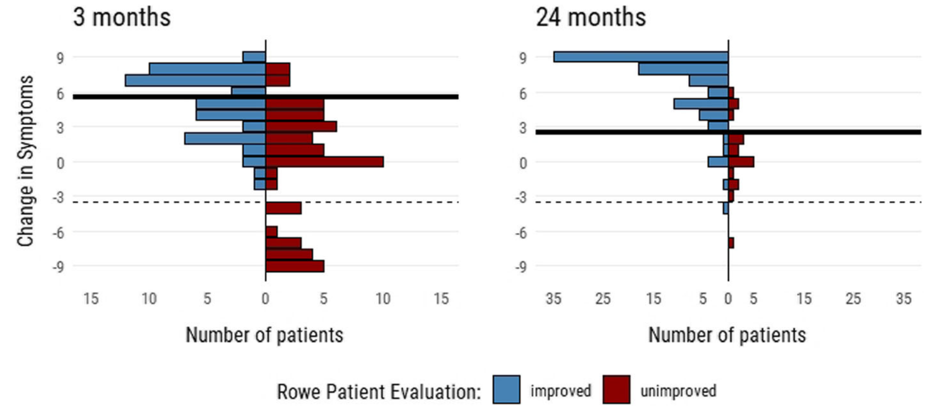

Fig. 1 Distribution of Change in Symptoms grouped by improved/unimproved at 3- and 24-months follow-up. Improved/unimproved is defined by Rowe Patient Evaluation were patients responding “Excellent” or “Good” are considered improved while patients responding “Fair” or “Poor” are considered unimproved. The solid black line marks the cut-off point for improved/unchanged. The dotted black line marks the cut-off between unchanged/deteriorated

RCP was defined as the percentage of patients improved by more than the MDC (Minimal Detectable Change). MDC estimates for all outcome measures, except CMS, were obtained from an earlier study using partially the same sample [7, 8]. MDC for the CMS was estimated by averaging the findings of earlier studies [23–26]. A robustness check for the CMS was performed using the lowest published MDC we could find [23]. 95% confidence intervals were estimated using the Clopper-Pearson method [27].

ROC analysis was incorporated to assess each instrument's ability to correctly classify patients as improved or unchanged. The sensitivity is defined as the proportion of improved patients correctly classified as improved, while the specificity is the proportion of unchanged patients correctly classified as unchanged. The ROC graph is a plot of the sensitivity against 1-specificity, illustrating the trade-off between false positives and true positives at all thresholds [28]. The optimal threshold was selected at the point on the ROC curve that minimizes the sum of squares of 1-sensitivity and 1-specificity, or equivalently the point closest to the upper-left corner [29].

Minimal important difference

MID is defined and estimated in a variety of ways in the literature, and there is a lack of formal agreement on which methods are superior [2]. This study incorporates anchor-based distribution methods (SRM, RCP, $\mathrm{MID}_{\mathrm{MEAN}}$) and ROC analysis (ROC cut-off, $\mathrm{MID}_{95\%\mathrm{limit}}$). ROC cut-off (often referred to as $\mathrm{MID}_{\mathrm{ROC}}$) is defined as the optimal threshold retrieved from ROC analysis [22, 30, 31]. 95% CI was estimated using a stratified bootstrap procedure, keeping the proportion of improved/ unchanged patients constant in each replicate sample. 95% CI for $\mathrm{ROC}_{\mathrm{AUC}}$ was estimated using the DeLong method [32].

$\mathrm{MID}_{\mathrm{MEAN}}$ was defined as the MCS of the patients scoring slightly above the chosen cut-off value on the anchor (i.e. for 6 months this equals patients scoring 4 and 5). The idea is that this group of patients consider themselves minimally improved, and their MCS can therefore be used to identify a MID estimate [30].

$\mathrm{MID}_{95\%\mathrm{limit}}$ was calculated as $\mu_{change} + 1.645 \cdot \sigma_{change}$ of the unchanged group. This corresponds to the 95% upper limit of the distributions of patients not experiencing an important improvement, and is equivalent to the cut-off value at the 95% specificity on the ROC curve [22]. Statistical analyses were performed in R, version 3.6.2 [33]. ROC analysis was performed using the pROC-package [34].

Hypotheses

Hypotheses were defined to further evaluate the responsiveness and validity of the instruments and anchor. The following null hypotheses were formulated for all instruments at all follow-ups if not otherwise stated:

1. MCS for men and women are equal.
2. MCS for patients above and below 40 years are equal.
3. The correlation between the MCS and change in symptoms (− 9 to 9) ≥ 0.70.
4. The correlation between the MCS and the anchor (improved/unchanged) ≥ 0.50.
5. MCS for patients with postoperative stiffness is ≤ the MCS for patients without postoperative stiffness at both 3- and at 6-months follow-up.
6. The correlation of the MCS between the instruments ≥0.70.

Hypotheses (1), (2) and (5) were tested using an independent sample t-test or Wilcoxon rank-sum test. Postoperative stiffness was defined as a loss of passive (glenohumeral) range of motion of > 30° in external rotation and abduction. Hypotheses (3) and (4) were tested using Spearman's rank correlation. Hypothesis (6) was tested using Pearson's r correlation. A correlation was defined as high > 0.70, as moderate between 0.40 and 0.70, and low < 0.40.

Results

119 patients were included (Table 2). Eight, four, six and six patients, respectively, did not answer the change in symptoms question at the different follow-ups and were excluded from the analysis. Sixteen (14%), 12 (10%), six (5%) and two (2%) patients answered below − 3 on the anchor at the different follow-ups and were considered deteriorated. The unchanged/improved ratio at 3, 6, 12-

Table 2 Descriptive statistics

Males/Females	**72/47**
Age (years), mean (range)	40.1 (18–64)
Duration of symptoms (months), median (range)	26 (6–360)
Dominant shoulder	89 (75)
Manual labour	47 (39)
Postoperative capsular stiffness	
3 months	30 (25)
6 months	17 (14)
Physical activity	
None	50 (42)
Weekly	61 (51)
Competition	7 (6)
Treatment	
Sham surgery	40 (34)
Biceps tenodesis	39 (33)
Labral repair	40 (34)

The number (%) of patients is given if not otherwise stated and 24-months follow-up was 64/31, 32/71, 17/90 and 21/90.

Baseline total score, MCS, SRM and RCP with 95% CI grouped by the anchor are reported in Table 3. In the improved group the SRM values for Rowe and WOSI were significantly higher than for the CMS and EQ-5D3L at all follow-ups, and for OISS at 3 months follow-up. The CMS was significantly higher than EQ-5D3L at 12- and 24-months follow-up. SRM values for EQ-5D3L ranged from 0.67 to 1.05 (moderate to high). All other SRM values were considered high in the improved group. In the unchanged group all SRM values were low at 3 months follow-up, ranged from low to moderate at 6- and 12-months follow-up, and moderate to high at 24 months follow-up.

RCP values in the improved group were highest for Rowe and WOSI at all follow-ups (ranging from 68 to 87%), with similar values for OISS at 6, 12- and 24-months. EQ-5D3L had the lowest values, and contrary to the other instruments, there was no increase over time (ranging from 13 to 16%). RCP values for the CMS ranged from 23 to 49%, using 16 as the MDC. When using 12 as the MDC the RCP values for the improved group were 32, 43, 56 and 69%.

ROC analysis and MID

Fifteen out of twenty ROC_{AUC} values were > 0.70 (Table 4). EQ-5D3L had the lowest values ranging from 0.55 to 0.74. ROC curves for all instruments at all follow-ups are reported in Fig. 2. Cut-off values were low for all scores at 3 months follow-up. The OISS, CMS and EQ-5D3L had confidence intervals crossing zero. At 6 months follow-up all cut-off values increased substantially, and were stable between 6- and 24-months follow-up. At 24 months follow-up ROC cut-off values (0 to 100 scale) was 18 for Rowe, 331 (16) for WOSI, 9 (19) for OISS, 11 for CMS and 0.123 (45) for EQ-5D3L. Excluding EQ-5D3L the largest difference between the instruments was 8 on a 0 to 100 scale.

$MID_{95\%limit}$ estimates were substantially higher than ROC cut-off values at all follow-ups for all instruments. The estimates peaked at 12 months follow-up for all instruments, while being comparable at the other follow-ups. At 24 months follow-up $MID_{95\%limit}$ values (0 to 100 scale) were 29 for Rowe, 853 (41) for WOSI, 16 (36) for OISS, 22 for CMS and 0.273 (54) for EQ-5D3L.

MID_{MEAN} values were higher than ROC cut-off values, but lower than $MID_{95\%limit}$ values, at 3 months follow-up for all instruments. At all other follow-ups MID_{MEAN} and ROC cut-off values were comparable. MID_{MEAN} values (0 to 100 scale) at 24 months follow-ups were 17 for Rowe, 401 (19) for WOSI, 10 (21) for OISS, 11 for CMS and 0.128 (45) for EQ-5D3L.

At 24 months follow-up MID estimates were lower than the MDC for CMS (except $MID_{95\%limit}$) and EQ-5D3L. For the other instruments all MID estimates were higher than the MDC (ROC cut-off value for WOSI was approximately equal).

Hypothesis testing

There was no evidence of any difference in mean change score between males and females or between patients aged below or above 40 years for any instrument at any follow-up (H1 and H2). All correlations between the MCS and the change in symptoms question were positive (the lowest being 0.33), but only two were > 0.70 (OISS at 6- and 12-months follow-up). The correlations ranged from 0.58 to 0.62 for Rowe, 0.55 to 0.69 for WOSI, 0.47 to 0.74 for OISS, 0.52 to 0.60 for CMS and 0.33 to 0.46 for EQ-5D. Correlations were lowest at 3 months follow-up (H3). Nine of sixteen correlations between the MCS and the anchor (improved/unchanged) for all instruments except EQ-5D were > 0.50. Correlations ranged from 0.43 to 0.60 for Rowe, 0.44 to 0.55 for WOSI, 0.40 to 0.59 for OISS, 0.38 to 0.56 for CMS and 0.13 to 0.32 for EQ-5D (H4). The MCS for patients with postoperative stiffness were significantly smaller than for patients without postoperative stiffness at 3 months follow-up for all instruments. At 6 months follow-up the null hypothesis could only be formally rejected for the Rowe score and the CMS (the *p*-value for OISS and WOSI was 0.05 and 0.06, respectively) (H5). The correlation of the MCS among the instruments at the different follow-ups ranged from 0.58 to 0.71 for Rowe/WOSI; 0.62 to 0.70 for Rowe/OISS; 0.74 to 0.81 for Rowe/CMS; 0.35 to 0.58 for Rowe/EQ-5D; 0.69 to 0.83 for WOSI/OISS; 0.59 to 0.71 for WOSI/CMS; 0.48 to 0.57 for WOSI/EQ-5D; 0.59 to 0.64 for OISS/CMS; 0.45 to 0.48 for OISS/EQ-5D and 0.35 to 0.48 for CMS/EQ-5D (H6).

Discussion

This study has evaluated responsiveness in five different outcome measures at four follow-ups. The MID estimates derived from the ROC analysis should be interpreted along with other estimates, particularly the mean change score and the measurement error of the instruments.

Estimates obtained at 3 months were not interpreted as meaningful for measuring outcome, particularly the confidence intervals for ROC cut-off values crossed zero for the OISS, CMS and the EQ-5D3L. This may reflect the short time period after surgery with large variability in the improvement process. Few patients had improved and some were deteriorated because of complications like stiff shoulders. Many patients had not regained their muscle strength. These factors influenced the scoring of outcome questions. A previous study evaluating patients

Table 3 Distribution based responsiveness in improved and unchanged patients

3 months										
Improved						Unchanged				
	Rowe	WOSI	OISS	CM	EQ-5D	Rowe	WOSI	OISS	CM	EQ-5D
Baseline	60	986	35	61.5	0.682	62.7	1013	34.3	63.3	0.649
0 to 100		47	48		79.5		48.2	46.5		77.3
95% CI	57 to 63	876 to 1096	32.6 to 37.5	57.8 to 65.2	0.612 to 0.741	60.1 to 65.3	916 to 1110	32.5 to 36.2	61 to 65.6	0.59 to 0.703
MCS	18.2	639	5.5	9.1	0.144	**1.1**	139	**0**	**9.0**	0.03
95% CI	14.3 to 22.1	506 to 771	3.3 to 7.6	4.9 to 13.2	0.075 to 0.221	**4.2** to 2.2	41 to 237	**1.8** to 1.8	**12.4** to **5.6**	**0.063** to 0.069
SRM	1.64	1.73	0.90	0.77	0.67	**0.08**	0.35	**0**	**0.65**	0.01
95% CI	1.19 to 2.39	1.41 to 2.28	0.58 to 1.34	0.48 to 1.16	0.47 to 0.93	**0.34** to 0.16	0.11 to 0.62	**0.25** to 0.25	**0.88** to **0.45**	**0.25** to 0.26
RCP (%)	68	70	26	23	13	13	32	10	0	8
95% CI	49 - 83	51 - 85	12 - 45	10 - 41	4 - 30	6 - 23	21- 45	4 - 20	0 - 6	3 - 18
6 months										
Improved						Unchanged				
	Rowe	WOSI	OISS	CM	EQ-5D	Rowe	WOSI	OISS	CM	EQ-5D
Baseline	61.9	1023	34.3	61.8	0.673	60.8	919	34.2	61.8	0.608
0 to 100		48.7	46.6		75.4		43.8	46.2		79.5
95% CI	59.5 to 64.3	941 to 1105	32.7 to 36	59.5 to 64.2	0.624 to 0.719	57.3 to 64.4	791 to 1047	31.6 to 36.8	58.9 to 64.8	0.515 to 0.693
MCS	22.4	690	13	11.7	0.163	3	232	3.5	**2.1**	0.105
95% CI	19.0 to 25.8	584 to 796	10.9 to 15	8.6 to 14.8	0.108 to 0.218	**1.0** to 7	120 to 344	2.0 to 5	**7.3** to 2.5	0.036 to 0.184
SRM	1.52	1.54	1.47	0.88	0.68	0.26	0.73	0.81	**0.15**	0.48
95% CI	1.11 to 2.18	1.26 to 1.92	1.17 to 1.9	0.61 to 1.23	0.42 to 0.96	**0.08** to 0.7	0.34 to 1.35	0.50 to 1.24	**0.46** to 0.24	0.25 to 0.71
RCP (%)	72	74	75	33	14	13	39	9	9	13
95% CI	60 - 82	62 - 84	64 - 85	22 - 45	7 - 25	4 - 29	22 - 58	2 - 25	2 - 25	4 - 29
1 year										
Improved						Unchanged				
	Rowe	WOSI	OISS	CM	EQ-5D	Rowe	WOSI	OISS	CM	EQ-5D
Baseline	62.5	1047	34.9	63	0.68	61.6	902	34.2	60.4	0.563
0 to 100		49.8	47.8		79.9		42.9	46.3		72.6
95% CI	60.3 to 64.6	973 to 1120	33.4 to 36.5	60.9 to 65	0.639 to 0.718	56.9 to 66.4	702 to 1102	30.2 to 38.2	57.3 to 63.4	0.423 to 0.689
MCS	25.1	742	15.5	13.7	0.174	8.2	251	4	6.6	0.153
95% CI	22.0 to 28.2	660 to 824	13.7 to 17.4	11.3 to 16.1	0.127 to 0.223	**0.1** to 16.5	40 to 473	0.1 to 8.1	2.0 to 11.3	0.018 to 0.288
SRM	1.67	1.88	1.76	1.17	0.75	0.45	0.53	0.46	0.67	0.52
95% CI	1.35 to 2.09	1.59 to 2.27	1.48 to 2.12	0.95 to 1.46	0.53 to 1.01	0 to 1.09	0.1 to 1.13	0.02 to 0.97	0.21 to 1.43	0.06 to 1.08
RCP (%)	76	87	80	32	13	35	35	29	6	24
95% CI	65 - 84	78 - 93	70 - 88	23 - 43	7 - 22	14 - 62	14 - 62	10 - 56	0 - 29	7 - 50
2 year										
Improved						Unchanged				
	Rowe	WOSI	OISS	CM	EQ-5D	Rowe	WOSI	OISS	CM	EQ-5D
Baseline	61.9	1024	34.5	62.4	0.673	62.8	979	35.9	62.0	0.62
0 to 100		48.8	46.9		79.5		46.6	49.7		76.2
95% CI	59.8 to 64	949 to 1099	33.0 to 36.1	60.4 to 64.4	0.629 to 0.712	59.4 to 66.1	820 to 1137	32.6 to 39.1	59.4 to 64.6	0.510 to 0.709
MCS	28.7	772	17.7	16.8	0.225	7.9	327	3.90	6.8	0.053
95% CI	25.9 to 31.5	693 to 852	15.9 to 19.4	14.5 to 19.1	0.181 to 0.270	2.3 to 13.5	197 to 471	0.5 to 7.3	2.7 to 10.9	**0.006** to 0.108
SRM	2.10	2.01	2.08	1.49	1.05	0.61	1.02	0.51	0.73	0.4
95% CI	1.72 to 2.62	1.71 to 2.43	1.79 to 2.47	1.25 to 1.82	0.87 to 1.28	0.2 to 1.14	0.69 1.6	**0.09** to 1.06	0.21 to 1.74	**0.05** to 1.02
RCP (%)	86	87	84	49	16	38	25	25	15	0
95% CI	77 - 92	78 - 93	75 - 91	38 - 60	9 - 25	18 - 62	9 - 49	9 - 49	3 - 38	0 - 17

Numbers in bold/red are negative values. *Baseline*, mean total score at baseline; *0 to 100,* baseline score transformed from original scale to a 0 to 100 scale; *MCS*, mean change score from follow-up to baseline; *SRM*, mean change score divided by the standard deviation of the mean change score; *RCP*, proportion of patients with change score exciding the Minimal Detectable Change

Table 4 Minimal important difference

3 months						
	ROC_{AUC} (95% CI)	ROC cut-off (95% CI)	Sens/Spec	MDC	MID_{MEAN}	$MID_{95\%limit}$
Rowe	0.80 (0.70-0.90)	8 (5-14)	0.79/0.78	14.3	14	20
WOSI	0.75 (0.64-0.86)	262 (206-530)	0.79/0.69	339.3	559 (27)	785 (37)
OISS	0.66 (0.54-0.77)	3 (**2**-5)	0.62/0.68	8.1	4 (8)	12 (25)
CM	0.78 (0.68-0.89)	0 (**2**-5)	0.76/0.76	16	9	14
EQ-5D	0.62 (0.51-0.74)	0.027 (**0.053**-0.172)	0.56/0.66	0.4	0.076 (42)	0.447 (65)

6 months						
	ROC_{AUC} (95% CI)	ROC cut-off (95% CI)	Sens/Spec	MDC	MID_{MEAN}	$MID_{95\%limit}$
Rowe	0.87 (0.80-0.94)	13 (6-17)	0.79/0.84	14.3	16	22
WOSI	0.79 (0.70-0.87)	450 (395-653)	0.71/0.81	339.3	354 (17)	754 (36)
OISS	0.83 (0.75-0.91)	8 (5-9)	0.77/0.91	8.1	6 (13)	11 (23)
CM	0.78 (0.69-0.88)	6 (4-8)	0.73/0.84	16	6	22
EQ-5D	0.63 (0.51-0.75)	0.036 (0.018-0.167)	0.69/0.66	0.4	0.089 (43)	0.467 (67)

12 months						
	ROC_{AUC} (95% CI)	ROC cut-off (95% CI)	Sens/Spec	MDC	MID_{MEAN}	$MID_{95\%limit}$
Rowe	0.77 (0.64-0.89)	12 (4-22)	0.77/0.71	14.3	17	38
WOSI	0.79 (0.66-0.92)	504 (202-514)	0.81/0.76	339.3	310 (15)	1028 (49)
OISS	0.82 (0.71-0.94)	9 (3-10)	0.79/0.76	8.1	6 (13)	18 (38)
CM	0.68 (0.55-0.82)	8 (4-13)	0.68/0.71	16	7	23
EQ-5D	0.55 (0.38-0.71)	0.079 (0.018-0.222)	0.64/0.59	0.4	0.050 (40)	0.639 (77)

24 months						
	ROC_{AUC} (95% CI)	ROC cut-off (95% CI)	Sens/Spec	MDC	MID_{MEAN}	$MID_{95\%limit}$
Rowe	0.86 (0.78-0.93)	18 (13 - 24)	0.77/0.81	14.3	17	29
WOSI	0.82 (0.71-0.92)	331 (289-442)	0.86/0.80	339.3	401 (19)	853 (41)
OISS	0.88 (0.80-0.96)	9 (5-14)	0.84/0.80	8.1	10 (21)	16 (36)
CM	0.75 (0.64-0.85)	11 (9-15)	0.58/0.71	16	11	22
EQ-5D	0.74 (0.64-0.85)	0.123 (0.035-0.222)	0.68/0.79	0.4	0.128 (45)	0.273 (54)

Numbers in bold/red are negative values. *ROC_{AUC}*, area under the ROC curve; *ROC cut-off*, change score threshold that minimizes the sum of squares of 1-sensitivity and 1-specificity (i.e. the point on the ROC curve closest to the upper-left corner); *Sens/Spec*, proportion of improved patients correctly classified as improved/proportion of unchanged patients correctly classified as unchanged; *MDC*, Minimal Detectable Change; *MID_{MEAN}*, mean change score of patients scoring slightly above the chosen cut-off value on the anchor (0 to 100 scale); *$MID_{95\%limit}$*, 95% upper limit of the change score distribution of patients defined as unchanged (0 to 100 scale)

with rotator cuff tears reported a MID of 10.4 at 3 months [35]. They did not report 95% CI and the ROC cut-off value was 2, which question their findings. We do not recommend the follow-up at 3 months after surgery for estimation of MID.

The distribution-based methods indicate that the CMS is less sensitive to change compared to OISS, WOSI and the Rowe score. MCS, SRM and RCP values were lower for CMS at all follow-ups. A different sensitivity to change was shown for the two clinical scores although baseline values for the improved group were similar. At 2 years the mean score for the CMS was 79.2 in the improved group, while it was 90.6 for the Rowe score. This indicates that for these patients the clinical scores do not scale equally. An increase of one point on the CMS equals a greater improvement on the Rowe score. This

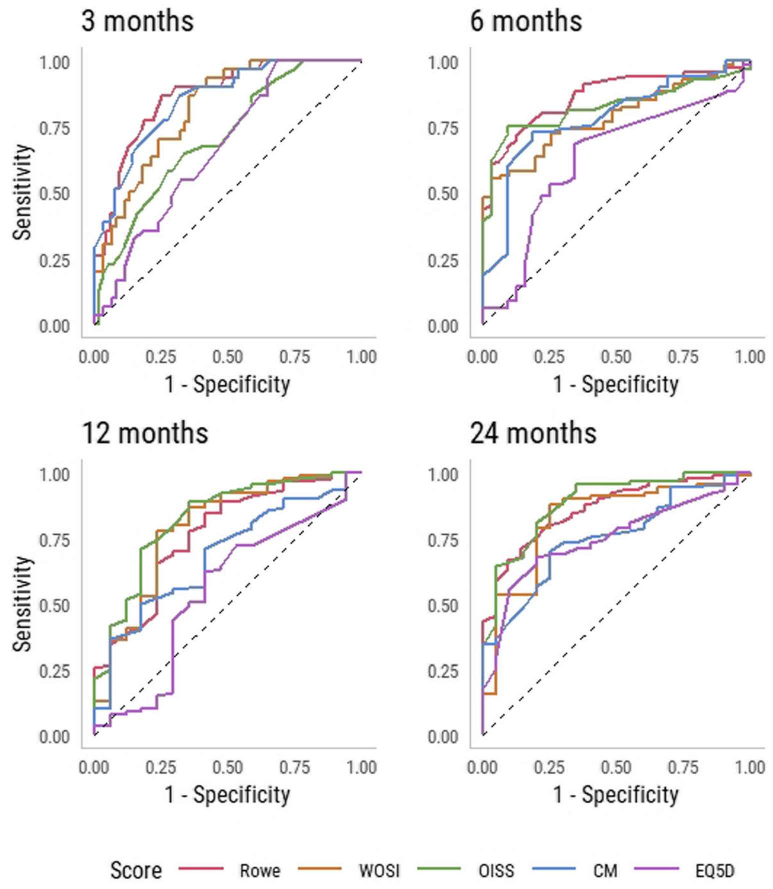

Fig. 2 ROC curves of all instruments at all follow-ups

might also explain why all MID estimates are lower for CMS than for the Rowe score and suggests that low MID values should not automatically be interpreted as better than higher ones.

Estimates of minimal important difference

The $MID_{95\%limit}$ estimates were substantially higher than the other MID estimates for all outcome measures. As de Vet et al. points out a challenging question is which cut-off point to prefer [22]. A factor driving the high $MID_{95\%limit}$ values was the high variation in change score among the unchanged patients. This possibly reflects that other health related issues affect their change score, or difficulties with the anchor in identifying patients who are minimally improved. Because every point on the ROC curve represents a trade-off between sensitivity and specificity, increasing the specificity to 0.95 comes at a cost. By example, at the 24 months follow-up the $MID_{95\%limit}$ for the Rowe score was 29 while the cut-off value when maximizing both was 18. The sensitivity and specificity were 0.77 and 0.81 for the latter (Table 4) and 0.56 and 0.95 for the former. We found no reason to dislike false negatives more than false positives, and therefore preferred the ROC cut-off value over the $MID_{95\%limit}$.

MID_{MEAN} is a less common way to measure the minimal clinically important change utilizing the fact that a continuous variable identifying the change in main complaint was collected at all follow-ups (− 9 to 9 scale). The disadvantage is related to identifying the group that one considers minimally improved. For example, at 6 months we defined patients scoring 4 and 5 as minimally improved and measured MID_{MEAN} as the mean change score among these patients. Alternatively, we could have used the median or the first or the third quartile emphasizing either *important* or *minimal* change. MID_{MEAN} values were comparable to ROC cut-off values at 6, 12- and 24-months follow-up.

At 24 months follow-up all MID estimates were lower than the MDC for the CMS and EQ-5D3L. In agreement with a recent systematic review we consider MID values that are lower than the measurement error (MDC) as problematic [2]. Recent studies present MID values that are below the MDC without any further discussion [2]. A disadvantage of the present study is that we have not provided MDC estimates for the CMS calculated in this sample. However, we conducted a robustness check using the lowest estimate published and the average MDC values from other studies.

The responsiveness of the EQ-5D3L was inferior to the other outcome measures. We consider this observation to be important because EQ-5D3L was recently used as an utility index in a systematic review to evaluate cost-effectiveness [6].

Hypothesis testing

We found a moderate correlation between mean change scores of the outcome measures and the anchor, except for the EQ-5D3L [36]. Low correlations between EQ-5D3L and the anchor are also illustrated by low ROC_{AUC} values, and the outcome measure is not suitable to detect improvements in this population.

Strengths and limitations

The strengths of this study include the use of both clinical scores and PROMS, at four different follow-ups. All clinical assessments were conducted by one single experienced assessor. Comprehensive statistical analyses were conducted for each outcome measure at all follow-ups. Estimates of MID were compared with estimates of MDC. The main challenge was identifying a valid anchor [1]. We found that some patients answered the anchor inconsistently, which may relate to the questionnaire or the heterogeneity of the patients. Some patients had complaints mainly related to specific sports, while others had daily pain and disability related to ordinary activities. Also, recall bias influence the anchor. A previous study found that global perceived change was influenced by the patients' state at the time of asking [37].

We recommend to use the MCS of the instruments as the primary outcome in trials rather than the proportion of patients exceeding the MID. On the other hand, MID values are helpful in calculating sample size and for understanding results in clinical practice but should be interpreted with an understanding of uncertainty and measurement error of the Instrument [7, 8, 38], the patient group and follow-up time examined.

Conclusions

In patients with SLAP II-lesions the patient reported OISS and WOSI and the clinical Rowe score had best responsiveness. Our results suggest that 3 months follow-up is too early for outcome evaluation. EQ-5D3L did not have appropriate measurement properties to assess responsiveness in this patient group.

Acknowledgments
We would like to thank Lovisenberg Diaconal Hospital, Lars Vasli, Chief of the Surgical Department, all the patients, the referring physicians, the nurses assisting in the operating theatre, the physiotherapists and manual therapists.

Authors' contributions
ØS and JIB designed the study. ØS, JIB and JSB participating in setting up the study. ØS and CPS recruited patients and collected data. ØS, JIB. and JSB wrote the main manuscript. JSB conducted the statistical analysis and prepared figures and tables. All authors reviewed the manuscript and gave their final approval of the submitted version.

Authors' information
Not applicable.

Funding
Lovisenberg Diaconal Hospital funded the study.

Availability of data and materials
The datasets generated and analyzed during the current study are not publicly available due to lack of patient consent and recommendation from University of Oslo. However, data can be provided on request to the corresponding author: oystein.skare@lds.no

Declarations

Ethics approval and consent to participate
The study was conducted according to the Declaration of Helsinki. Ethics approval (IRB00001870) was received from the Ethics Committee Health Region Southeast, Oslo, Norway. We obtained written informed consent from all participants.
The protocol was registered at ClinicalTrials.gov (NCT00586742).

Consent for publication
Not applicable.

Competing interests
None declared.

Author details
[1]Surgical Department, Lovisenberg Diaconal Hospital, Oslo, Norway. [2]Orthopedic Department, Lovisenberg Diaconal Hospital, Lovisenberggt 17, 0440 Oslo, Norway. [3]Copenhagen Business School, Copenhagen, Denmark. [4]Department of Physical Medicine and Rehabilitation, Oslo University Hospital, Oslo, Norway. [5]Medical Faculty, University of Oslo, Oslo, Norway.

Received: 5 March 2021 Accepted: 15 June 2021
Published online: 05 July 2021

References
1. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ. 2020;369:m1714.
2. Copay AG, Chung AS, Eyberg B, Olmscheid N, Chutkan N, Spangehl MJ. Minimum clinically important difference: current trends in the Orthopaedic literature, part I: upper extremity: a systematic review. JBJS Rev. 2018;6(9):e1. https://doi.org/10.2106/JBJS.RVW.17.00159.
3. Feinstein AR. T. Duckett Jones memorial lecture. The Jones criteria and the challenges of clinimetrics. Circulation. 1982;66(1):1–5. https://doi.org/10.1161/01.CIR.66.1.1.
4. Fava GA, Tomba E, Sonino N. Clinimetrics: the science of clinical measurements. Int J Clin Pract. 2012;66(1):11–5. https://doi.org/10.1111/j.1742-1241.2011.02825.x.
5. Brox JI. The fear avoidance beliefs questionnaire - the FABQ - for the benefit of another 70 million potential pain patients. Scand J Pain. 2019;19(1):1–2. https://doi.org/10.1515/sjpain-2018-2005.
6. Paoli AR, Gold HT, Mahure SA, Mai DH, Agten CA, Rokito AS, et al. Treatment for symptomatic SLAP tears in middle-aged patients comparing repair, biceps Tenodesis, and nonoperative approaches: a cost-effectiveness analysis. Arthroscopy. 2018;34(7):2019–29. https://doi.org/10.1016/j.arthro.2018.01.029.
7. Skare Ø, Liavaag S, Reikerås O, Mowinckel P, 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. BMCResNotes. 2013;6:273.

8. Skare Ø, Schrøder CP, Mowinckel P, Reikerås O, Brox JI. Reliability, agreement and validity of the 1988 version of the Rowe score. J Shoulder Elb Surg. 2011;20(7):1041–9. https://doi.org/10.1016/j.jse.2011.04.024.
9. Skare Ø, Mowinckel P, Schrøder CP, Liavaag S, Reikerås O, Brox JI. Responsiveness of outcome measures in patients with superior labral anterior and posterior lesions. Should Elb. 2014;6(4):262–72. https://doi.org/10.1177/1758573214534650.
10. Kemp KA, Sheps DM, Beaupre LA, Styles-Tripp F, Luciak-Corea C, Balyk R. An evaluation of the responsiveness and discriminant validity of shoulder questionnaires among patients receiving surgical correction of shoulder instability. Sci World J. 2012;2012:410125.
11. Schrøder CP, Skare Ø, Reikerås O, Mowinckel P, Brox JI. Sham surgery versus labral repair or biceps tenodesis for type II SLAP lesions of the shoulder: a three-armed randomised clinical trial. Br J Sports Med. 2017;51(24):1759–66. https://doi.org/10.1136/bjsports-2016-097098.
12. Brox JI, Skare Ø, Mowinckel P, Brox JS, Reikerås O, Schrøder CP. Sick leave and return to work after surgery for type II SLAP lesions of the shoulder: a secondary analysis of a randomised sham-controlled study. BMJ Open. 2020;10(4):e035259. https://doi.org/10.1136/bmjopen-2019-035259.
13. Skare Ø, Schrøder CP, Reikerås O, Mowinckel P, Brox JI. Efficacy of labral repair, biceps tenodesis, and diagnostic arthroscopy for SLAP lesions of the shoulder: a randomised controlled trial. BMC Musculoskelet Disord. 2010;11(1). https://doi.org/10.1186/1471-2474-11-228.
14. Kirkley A, Griffin S, McLintock H, 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(6):764–72. https://doi.org/10.1177/03635465980260060501.
15. Dawson J, Fitzpatrick R, Carr A. The assessment of shoulder instability. The development and validation of a questionnaire. J Bone Joint Surg (Br). 1999;81(3):420–6. https://doi.org/10.1302/0301-620X.81B3.0810420.
16. EuroQolGroup. EuroQol: A new facility for the measurement of health related quality of life. Health Policy. 1990;16:199–208. https://doi.org/10.1016/0168-8510(90)90421-9.
17. Constant CR, Murley AHG. A clinical method of functional assessment of the shoulder. Clin Ortop. 1987;214:160–4.
18. Rowe CR, Partel D, Sothmayd WW. The Bankart procedure; a long-term and end-result study. J Bone Joint Surg Am. 1978;60-A:1–16.
19. Brox JI, Gjengedal E, Uppheim G, Bohmer AS, Brevik JI, Ljunggren AE, et al. Arthroscopic surgery versus supervised exercises in patients with rotator cuff disease (stage II impingement syndrome): a prospective, randomized, controlled study in 125 patients with a 2 1/2-year follow-up. J Shoulder Elb Surg. 1999;8(2):102–11.
20. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012.
21. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147–57. https://doi.org/10.1007/s11136-018-1798-3.
22. de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res. 2007;16(1):131–42. https://doi.org/10.1007/s11136-006-9109-9.
23. Møller AD, Thorsen RR, Torabi TP, Bjørkman AS, Christensen EH, Maribo T, et al. The Danish version of the modified Constant-Murley shoulder score: reliability, agreement, and construct validity. J Orthop Sports Phys Ther. 2014;44(5):336–40. https://doi.org/10.2519/jospt.2014.5008.
24. Mahabier KC, Den Hartog D, Theyskens N, Verhofstad MHJ, Van Lieshout EMM, Investigators HT. Reliability, validity, responsiveness, and minimal important change of the disabilities of the arm, shoulder and hand and Constant-Murley scores in patients with a humeral shaft fracture. J Shoulder Elb Surg. 2017;26(1):e1–e12. https://doi.org/10.1016/j.jse.2016.07.072.
25. Blonna D, Scelsi M, Marini E, Bellato E, Tellini A, Rossi R, et al. Can we improve the reliability of the Constant-Murley score? J Shoulder Elb Surg. 2012;21(1):4–12. https://doi.org/10.1016/j.jse.2011.07.014.
26. Henseler JF, Kolk A, van der Zwaal P, Nagels J, Vliet Vlieland TP, Nelissen RG. The minimal detectable change of the Constant score in impingement, full-thickness tears, and massive rotator cuff tears. J Shoulder Elb Surg. 2015;24(3):376–81. https://doi.org/10.1016/j.jse.2014.07.003.
27. Clopper CJ, Pearson ES. The Use of Confidence or Fidusial Limits Illustrated in the Case of Binomial. Biometrika. 1934;26(4):404–13.
28. Fawcett T. Introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–74. https://doi.org/10.1016/j.patrec.2005.10.010.
29. Froud R, Abel G. Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: the forgotten lesson of pythagoras. theoretical considerations and an example application of change in health status. PLoS One. 2014;9(12):e114468.
30. Ekeberg OM, Bautz-Holter E, Keller A, Tveita EK, Juel NG, 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(5):575–84. https://doi.org/10.1016/j.jclinepi.2009.07.012.
31. Holmgren T, Oberg B, Adolfsson L, Bjornsson Hallgren H, Johansson K. Minimal important changes in the Constant-Murley score in patients with subacromial pain. J Shoulder Elb Surg. 2014;23(8):1083–90. https://doi.org/10.1016/j.jse.2014.01.014.
32. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45. https://doi.org/10.2307/2531595.
33. Team RC. A language and envronment for statistical computing. In: R package version 3.6.1, 2019. Vienna, Austria: R Foundation for Statistical Computing; 2019.
34. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform. 2011;12(1):77. https://doi.org/10.1186/1471-2105-12-77.
35. Kukkonen J, Kauko T, Vahlberg T, Joukainen A, Aarimaa V. Investigating minimal clinically important difference for Constant score in patients undergoing rotator cuff surgery. J Shoulder Elb Surg. 2013;22(12):1650–5. https://doi.org/10.1016/j.jse.2013.05.002.
36. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol. 2002;55(9):900–8. https://doi.org/10.1016/S0895-4356(02)00435-3.
37. Grøvle L, Haugen AJ, Hasvik E, Natvig B, Brox JI, Grotle M. Patients' ratings of global perceived change during 2 years were strongly influenced by the current health status. J Clin Epidemiol. 2014;67(5):508–15. https://doi.org/10.1016/j.jclinepi.2013.12.001.
38. Conboy VB, Morris RW, Kiss J, Carr AJ. An evaluation of the Constant-Murley shoulder assessment. J Bone Joint Surg (Br). 1996;78(2):229–32.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.