Study design and outcome measures for evaluation of the treatment for SLAP lesions in the shoulder

Thesis

Øystein Skare

Orthopaedic Department, Lovisenberg Diaconal Hospital

Orthopaedic Department, Oslo University Hospital, Rikshospitalet

University of Oslo, Institute of Clinical Medicine

2014
He who loves practice without theory is like the sailor who boards ship without a rudder and a compass and never knows where he may cast.

Leonardo da Vinci
# Table of contents

Acknowledgements ........................................................................................................ 4
Thesis at a glance .......................................................................................................... 5
List of papers ................................................................................................................ 7
Abbreviations ................................................................................................................. 8
1 General introduction ................................................................................................. 9
   1.1 Epidemiology of shoulder pain .......................................................................... 9
   1.2 Superior labral anterior to posterior (SLAP) lesions ........................................ 9
2. Aims of the thesis ..................................................................................................... 17
3. Patients and methods ............................................................................................. 18
   3.1 Design ............................................................................................................... 18
   3.2 Patients ............................................................................................................. 18
   3.3 Procedures and measurements ...................................................................... 20
   3.4 Statistics .......................................................................................................... 23
   3.5 Ethical issues .................................................................................................. 30
4. Synopsis .................................................................................................................. 32
   4.1 Paper I ............................................................................................................ 32
   4.2 Paper II ........................................................................................................... 33
   4.3 Paper III .......................................................................................................... 35
   4.4 Paper IV .......................................................................................................... 36
5. General discussion ................................................................................................. 38
   5.1 Patient sample and external validity ................................................................. 38
   5.2 Patient sample and internal validity ................................................................. 38
   5.3 Study designs .................................................................................................. 39
   5.4 Statistical approach ......................................................................................... 41
   5.5 Outcome measures ......................................................................................... 44
   5.6 Results ............................................................................................................ 45
6. General conclusions ............................................................................................... 49
   Paper I-III .............................................................................................................. 49
   General conclusion ............................................................................................... 50
References ..................................................................................................................... 51
   Paper I-IV .............................................................................................................. 59
Acknowledgements

This work was carried out at the department of orthopaedic surgery at Lovisenberg Diaconal Hospital (2006-2014). Funding was provided by Lovisenberg Diaconal Hospital.

I want to express my sincere gratitude to the main supervisor professor Jens Ivar Brox and co-supervisor professor Olav Reikerås for their excellent supervision. Their experience from research, from writing scientific publications, and their ability to put the methodological details into a perspective has been of the greatest value. Professor Jens Ivar Brox has challenged me to solve problems and reach “the next top of knowledge” on my own, but have always been there to guide me and correct me in the end. I will especially thank him for the opportunity to learn this field of science.

The director of the surgical department of Lovisenberg Diaconal Hospital, Lars Vasli deserves a great honour for the opportunity to do this research. I am privileged to have a leader who understand the importance of clinical related research and who give time and resources to make it possible. My dear friend, colleague and research-fellow Cecilie Piene Schröder has been a great support and companion through planning and accomplishment of the studies, and through advices in the writing of the papers. She has also been a great inspiration for me in the clinical work and has introduced me to the field of shoulder surgery and research. I will express my thankfulness to my dear colleagues Magnus Wallumrød, Kirsten Lundgreen, Sigbjørn Dimmen and Bjørn Gunnar Kristiansen for advice and support, and to all of the staff and my colleagues at surgical department at Lovisenberg Diaconal Hospital for encouraging my research.

Sigurd Liavåg at Sørlandet Hospital, Arendal has been a great support and inspiration trough the whole process. He has initialised the cross cultural adaptation of some of the measurement outcomes and has contributed to the writing of some of the papers. Statistician Petter Mowinckel has been a very important guide through this field of research consisting of so little consensus and so many methodological challenges. His ability to find the right solutions is admirable. Last, but certainly not least, I wish to thank my beloved family for their endurance, love and support.
## Thesis at a glance

<table>
<thead>
<tr>
<th>Aim</th>
<th>Method</th>
<th>Results</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>To evaluate reliability, agreement and validity of the 1988 version of the clinical Rowe Score on patients with SLAP lesions and recurrent dislocations using recommended statistical methods.</td>
<td>Seventy-one patients (33 with recurrent anterior dislocations and 38 with SLAP lesions) were tested by two experienced clinicians performing the Rowe score independently twice with 7 days interval. To allow comparison with other outcome measures in evaluation of construct validity by hypothesis method, patients completed the Oxford Instability Shoulder Score (OISS), Western Ontario Shoulder Instability Index (WOSI) and EuroQol.</td>
<td>No significant differences between clinicians and test days were detected for the Rowe score. Limits of agreement of the total score varied from 17.9 to 20.5. For the categorical score (total units), there were significant differences between the two examiners (P &lt; .001). The ICC 2.1 was acceptable (&gt; .70) for the total score. No floor or ceiling effects were observed for the total score, but considerable floor and/or ceiling effects were detected for some of the domains. Discriminant validity was acceptable, but content, construct, and convergent validity was not acceptable.</td>
</tr>
</tbody>
</table>

| 2 | To evaluate agreement, reliability and validity of two commonly used questionnaires developed for patients with shoulder instability, and a generic questionnaire in patients with type II SLAP lesions or recurrent anterior dislocations | Seventy-one patients (33 with recurrent anterior dislocations and 38 with SLAP lesions) completed the OISS, WOSI and EuroQol twice at the same time of day (±2 hours) with a one week interval between administrations. Hypothesis method was used to evaluate construct validity. | ICC ranged from 0.89 (95% CI 0.83 to 0.93) to 0.92 (0.87 to 0.95) for OISS, WOSI and EQ-VAS and was 0.66 (0.50 to 0.77) for EQ-5D. The limits of agreement for the scores were: -7.8 to 8.4 for OISS; -399.9 to 344.8 for WOSI; -0.4 to 0.4 for EQ-5D; and -17.2 to 16.2 for EQ-VAS. All questionnaires reflect the construct to that was measured. The correlation between WOSI and OISS was 0.73, and ranged from 0.49 to 0.54 between the shoulder questionnaires and the generic questionnaires. The divergent validity was acceptable, convergent validity failed, and known group validity was acceptable only for OISS. | Measurement errors and limitations should be considered when change scores of OISS and WOSI are interpreted in patients with SLAP lesions or recurrent anterior shoulder dislocations. EQ-5D is not recommended as a single outcome. |

To be continued
<table>
<thead>
<tr>
<th>Aim</th>
<th>Method</th>
<th>Results</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>To evaluate responsiveness and minimal clinical important change of the 1988 version of the clinical Rowe Score, the specific Oxford Instability Shoulder Score and Western Ontario instability Index, and the generic EuroQol used on patients with type 2 SLAP lesions.</td>
<td>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 six months. Responsiveness was evaluated using standardised response mean (SRM), area under receiver operating characteristic curve (ROC_{AUC}), reliable chance proportion (RCP) statistics, and hypothesis method. Minimal clinical important change (MCIC) estimates were reported.</td>
<td>All outcome measures had high values of SRM (0.86-1.92). RCP’s for the improved group were 68-79% for OISS, WOSI and Rowe score, and 15-49% for EuroQol. ROC_{AUC} was &gt; 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.</td>
</tr>
<tr>
<td>4</td>
<td>To compare the short-term (6 months) and long-term (2 years) efficacy of labral repair, biceps tenodesis, and placebo (diagnostic arthroscopy) for alleviating pain and improving function for type II SLAP lesions.</td>
<td>Protocol article: Design and performance of a prospective, randomised, double blinded, sham-controlled study with 120 patients. Three group design; 1) Arthroscopic repair of labrum 2) Mini-open biceps tenodesis 3) Diagnostic arthroscopy All groups: Standardised but individual adjusted postoperative rehabilitation</td>
<td>Inclusion is ended. Follow up is still running. The results are not analysed.</td>
</tr>
</tbody>
</table>
List of papers

Paper I
Reliability, agreement and validity of the 1988 version of the Rowe Score
Skare Ø, Schrøder CP, Mowinckel P, Reikerås O, Brox JI.

Paper II
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.
Skare Ø1, Liavaag S, Reikerås O, Mowinckel P, Brox JI

Paper III
Responsiveness of outcome measures in patients with superior labral anterior and posterior (SLAP) lesions
Skare Ø, Mowinckel P, Schrøder CP, Liavaag S, Reikerås O, Brox JI
Shoulder & Elbow published online 27 May 2014. doi: 10.1177/1758573214534650

Paper IV
STUDY PROTOCOL: Efficacy of labral repair, biceps tenodesis, and diagnostic arthroscopy for SLAP Lesions of the shoulder: a randomised controlled trial
Skare Ø, Schrøder CP, Reikerås O, Mowinckel P, Brox J
<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLAP lesions</td>
<td>Superior labral anterior and posterior lesions</td>
</tr>
<tr>
<td>OISS</td>
<td>Oxford Instability Shoulder Score</td>
</tr>
<tr>
<td>WOSI</td>
<td>Western Ontario Shoulder Instability Index</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>EuroQol 5 Domains</td>
</tr>
<tr>
<td>EQ-VAS</td>
<td>EuroQol – visual analogue scale</td>
</tr>
<tr>
<td>ICC</td>
<td>Interclass correlation coefficient</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic resonance imaging</td>
</tr>
<tr>
<td>MCIC</td>
<td>Minimal clinical important change</td>
</tr>
<tr>
<td>MDC</td>
<td>Minimal detectable change</td>
</tr>
</tbody>
</table>
1 General introduction

1.1 Epidemiology of shoulder pain

Pain and disability in the shoulder have been estimated to be the third most frequent musculoskeletal complaint in general practice, next to complaints from the lower back and neck [1]. The point prevalence of shoulder pain in the general population range from 7 to 26 %, with a one-year prevalence between 5 and 47% [2;3]. In a Norwegian survey, the prevalence of weekly neck- and shoulder pain has been reported to be 13% for men and 25% for women [4]. A paper based on the Oslo Health Study in Norway (2000 – 2001), reported pain in the neck/shoulder in 52 % of the responders. Of these, 38% reported work-related neck/shoulder pain [5]. In a Norwegian study from 2013, investigating work-related pain in the neck and upper extremity, the threshold for reporting pain seemed higher and clinical diagnosis more frequent in subjects with low socioeconomic position and among women. [6].

The distribution of specific diagnosis related to pain in the shoulder is not well documented in the general population. Most studies of prevalence and incidence of specific diagnosis are obtained from patients seeking help for the shoulder pain. Clinical studies indicate that subacromial structures are most frequent affected and that rotator cuff disorders and subacromial impingement are the most frequently used specific shoulder diagnosis [7-9].

1.2 Superior labral anterior to posterior (SLAP) lesions

The first one to describe pathological findings of glenoid labrum was Olsson in 1953 [10]. With the entry of arthroscopic surgery of the shoulder in the 1980’s, the diagnostic possibilities and classifications of injuries decreased. In 1985, Andrews et al described lesions of the glenoid labrum in a material of 73 throwing athletes, who had an arthroscopic examination of their dominant shoulder. Most of the tears in their material occurred at the antero-superior portion of the glenoid labrum, near by the origin of the long head of the biceps tendon [11].
1.2.1 Classification

The term superior labrum anterior and posterior (SLAP) lesions were originally described by Snyder et al in 1990 [12]. The SLAP lesions were divided into four types. The Type I is described as a marked fraying with a degenerative appearance of the superior labrum, with the peripheral labral edge remained firmly attached to the glenoid, and the attachment of the biceps tendon intact (Figure 1). Type II lesion has similar fraying and degenerative changes as Type I, but with an additional loosening of the superior labrum from the glenoid, resulting in an unstable labral-biceps anchor arched away from the glenoid. Type III lesion is a bucket-handle tear of the superior labrum, with a displacement into the joint of the central portion of the tear. The peripheral portion of the labrum remains attached to the underlying glenoid and to the biceps tendon, which is also intact. The type IV lesion is similar to type II, but in addition, the tear includes the biceps tendon with an attached partial tear.

Figure 1 Arthroscopic and schematic illustration of SLAP lesion Type I (from Snyder et al 1990 [12])

Figure 2 Arthroscopic and schematic illustration of SLAP lesion Type II (from Snyder et al 1990 [12])
In 1995, Maffet and co-workers described further types of SLAP lesions, suggesting an expanding of Snyder’s classification with additional three types (Type V-VII) [13]. A classification of Type II SLAP lesions into three subtypes were described by Morgan et al in 1998. Type A anterior; Type B posterior; Type C combined anterior and posterior [14].
1.2.2 Incidence and prevalence

The true prevalence and incidence of SLAP lesions is still unknown. In arthroscopic studies, SLAP lesions have been found in 3.9% to 26% of the cases [12;15;16]. In the study of Snyder published in 1995 based on 2375 shoulder arthroscopies, 140 patients (6%) had SLAP lesions. Twenty-one percent were Type I, 55% were Type II, 9% were Type III, and 10% were Type IV. An isolated SLAP lesion was found in 28% of the patients. Associated lesions of partial thickness rotator cuff tears was seen in 29%, full thickness ruptures in 11%, and Bankart lesions in 22%. [16]. Morgan et al found in 1998 rotator cuff tears in 31% of patients with SLAP lesions in a material of 102 patients. They concluded that SLAP lesions with a posterior component were associated with the development of posterior–superior instability manifested as a secondary anterior-inferior pseudo laxity, and that chronic superior instability were associated with secondary lesion-location-specific rotator cuff tears that begins as partial thickness tears from inside the joint [14]. In 2003, Kim et al reported an occurrence of SLAP lesions at the level of 26% of 544 consecutive shoulder arthroscopies [15]. Seventy-four percent were Type I lesions, 21% were Type II lesions, 0.7% were Type II lesions, and 4% were Type IV lesions according to Snyder’s classification of SLAP lesions [15]. In a recent study of Weber et al, 9.4% of all applicants’ shoulder cases (4975) were SLAP repairs. The number of SLAP
repairs increased to 10 % during the study period (2003-2008). Of all SLAP repairs, 13 % of the patients had isolated SLAP lesions [17].

The increased accessibility of Magnetic Resonance Imagine (MRI) and MRI arthrography (injection of contrast fluid into the shoulder joint) have contributed to the diagnostic of SLAP lesions, but also of capsulolabral anatomic variants such as Buford complex, which can be a challenge to divide from pathological findings of a SLAP lesion.

### 1.2.3 Treatment of SLAP lesions

Before development of arthroscopy and MRI, symptoms and functional problems due to SLAP lesions were probably untreated or recognized as subacromial impingement, and treated with open acromioplasty or physiotherapy. The first specific arthroscopic treatment for SLAP lesions involved debridement of the labrum, but the long-term results were not promising [18;19]. As a result of this knowledge, reattachment of the labrum to the glenoid became increasingly more popular as the surgical treatment of SLAP lesions during the 1990’s and later. Different methods have been used for surgical repair including metal screws, staples, sutures, bioabsorable tacks, and suture anchors [19-22]. As an alternative to arthroscopic repair of type II SLAP lesions Boileau et al introduced arthroscopic biceps tenodesis, which means that the long head of the biceps were removed from the glenoid labrum (biceps tenotomy) and fixed to the top of the bicipital groove using an interference screw [23]. In a non-randomised cohort study with 25 patients, 10 patients received arthroscopic repair with resorbable suture anchors and 25 patients were treated with biceps tenodesis. Patient satisfaction and the ability to return to previous level of sports participation favoured biceps tenodesis [23]. There is little information about the effect of non-operative treatment of SLAP lesions. Such treatment have been described as stretching of the posterior capsular while maintaining periscapular and glenohumeral strength [24-26] and the use of nonsteroidal anti-inflammatory drugs together with scapular stabilization exercises and posterior capsular stretching [27]. No results of randomised controlled trials involving surgical or non-operative treatment of SLAP lesions have been published.
1.2.4 Outcome assessment

Shoulder function has traditionally been evaluated at an impairment level by clinicians measuring range of motion and strength, with or without instruments. In the 1970’s and 1980’s this measurements were combined with questions assessing pain and function [28-32]. During the last four decades patient reported health related quality of life has been increasingly more used in the evaluation of patients [33-40]. By this evolution, the perspective of judgment of treatment effects has been moved from a clinical evaluation to a patient’s perspective. Kirschner and Guyatt [41] have identified three main characteristics of methods used in evaluation of health related quality of life; 1) discriminative (to discriminate among individuals), 2) predictive (to predict future health status), and 3) evaluative (to detect important changes over time in health status). The third aspect responsiveness was highlighted in the present thesis. Measures of outcome should preferably have good measurement properties in terms of being reliable, valid and responsive to be useful in clinical practice and research [42;43]. The patient reported health related quality of life outcomes may be divided into three categories; disease specific, region specific, and generic. The specific outcomes are developed to assess quality of life related to the patients perceptions of quality of life related to the disease or region (shoulder pain), while the generic outcomes are designed to measure quality of life in general in a wide range of health problems.

1.2.5 Background for the current study

We had previously conducted a prospective study with 5 years to assess the effectiveness for capsular repair in patients [22]. The main outcome was the Rowe score evaluated by one independent observer. The mean improvement was 29.3 points (95% CI 26.1 to 32.4). Although patients included in this study reported to have complaint for 52 months (range 4 to 267), factors such as the natural course, regression to the mean, and placebo, may have contributed to improvement. Therefore, a prospective study without control group is in general considered to represent a low level of evidence [44]. The inclusion of a control group will improve the evidence, but groups may be different and bias results. Randomisation will usually create similar groups and reduce confounding. The double blind sham-
controlled trial is considered to represent the highest evidence achievable from a single clinical study.

(Figure from www.life-enhancement.com)

With this study design the sham group include similar procedures as the active intervention, by example a placebo injection is given to mimic an active injection and a surgical procedure, by example diagnostic arthroscopy is given to mimic all the procedures of the operation except the intended treatment (by example SLAP repair). After evaluation of the prospective study of SLAP repair we discussed several study designs to improve the scientific standard of the evaluation in a new study. The use of a randomised design to evaluate the effectiveness of shoulder surgery was first used in Norway in 1993 by Brox et al [45]. The trial was conducted in cooperation with shoulder surgeons at Lovisenberg Diaconal Hospital (at that time Menighetssøsterhjemmet Hospital) and physiotherapists at the Department of Physical Medicine and Rehabilitation at Ullevål Hospital. The trial included a placebo group and contributed to better understanding of the effectiveness of both surgery and physiotherapy in the treatment of the impingement syndrome. The placebo group had placebo laser and it could therefore not be excluded that the effects of surgery and physiotherapy was attributed expectations. To improve the study design we
therefore wanted to include sham surgery and finally agreed to plan a sham-controlled, double blind (patient and observer), randomised trial.

We also planned to use various questionnaires that were not applied in the prospective study. Besides using the Rowe score, we included several questionnaires to find the easiest and the best method to assess patients. By example, for evaluation a questionnaire including 10 questions is easier to handle than a more comprehensive questionnaire. A questionnaire can be mailed to the patients and is therefore most likely also cost saving compared to clinical assessment by a physiotherapist or physician. The ultimate goal is to use a questionnaire that address the patient’s problem (validity) with little measurement error (reliability) and is able to detect a change if there is a change (responsiveness). Additionally, the use of generic health –related quality-of-life outcomes are recommended in studies of comparative effectiveness to allow for comparison across patient populations [46;47]. To our knowledge, no outcome measures has been designed or validated for evaluation of the treatment of patients with SLAP lesions. The results of the sham-controlled randomised trial would be hampered without the use of outcome measures with acceptable psychometric properties like agreement, reliability, validity and responsiveness. We decided to evaluate one clinical score, two commonly used specific patient reported and one generic health-related quality-of-life outcomes for the purpose. The clinical and the specific patient reported outcomes were chosen because they had been used for assessment of shoulder instability, which probably is the diagnosis most related to SLAP lesions, and because they were the best we could find to reflect the complaints in the patients with SLAP lesions in domains like pain, function, and other aspects of quality of life. The generic outcome is commonly used in both comparisons across patient populations and for us in cost-effectiveness studies.
2. Aims of the thesis

The aims of this thesis were to evaluate outcome measures for use in evaluation of patients with SLAP lesions or recurrent anterior instability of the shoulder, and to plan and design a sham controlled, randomised, double blind study for evaluation of treatment efficacy for patients with SLAP lesions of the shoulder.

The particular aims of each study were:

1. To evaluate reliability, agreement, validity of the 1988 version of the clinical Rowe score (paper I)
2. To evaluate two commonly used questionnaires developed for patients with shoulder instability and a generic questionnaire (paper II) in patients with type II SLAP lesions or recurrent anterior dislocations.
3. To evaluate 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 (paper III).
4. To plan and conduct a randomised sham- controlled double- blinded study that compares the short-term (6 months) efficacy of labral repair, biceps tenodesis, and placebo (diagnostic arthroscopy), for alleviating pain and improving function for type II SLAP lesions (paper IV).
3 Patients and methods

3.1 Design

The thesis consists of three methodological studies and the planning of one trial. In the first study (paper I and II), a test-retest reliability design and a validation design were used. In the second study (paper III), a prospective longitudinal design was used. The third study was a randomized sham-controlled double-blind study design with two years follow up (paper IV).

3.2 Patients

Patients for all studies were referred to an outpatient clinic at department of orthopaedics at Lovisenberg Diaconal Hospital from physicians in primary care, outpatient clinics or accident and emergency units. All studies were approved by the ethical committee for medical research, and all patients attending the studies gave their informed consent.

3.2.1 Inclusion and exclusion criteria

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 16 to 18 years (paper I-II)</td>
<td></td>
</tr>
<tr>
<td>Age 18 to 60 years (paper II-IV)</td>
<td></td>
</tr>
<tr>
<td>Type II SLAP lesion based on anamnesis, primary arthroscopic findings, clinical findings, and MR arthrography</td>
<td></td>
</tr>
<tr>
<td>Type II SLAP lesion verified by arthroscopy (paper III-IV)</td>
<td></td>
</tr>
<tr>
<td>Recurrent anterior dislocations (paper I-II)</td>
<td></td>
</tr>
</tbody>
</table>
Exclusion criteria

<table>
<thead>
<tr>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous surgery for SLAP lesions (paper I-IV)</td>
</tr>
<tr>
<td>SLAP lesions with concomitant labral cysts (paper IV)</td>
</tr>
<tr>
<td>Previous surgery for recurrent shoulder dislocation, posterior or multidirectional dislocations (paper I to IV)</td>
</tr>
<tr>
<td>Pain referred from the cervical- or thoracic spine (paper I to IV)</td>
</tr>
<tr>
<td>Clinical and radiological signs of arthritis of the acromioclavicular joint or the glenohumeral joint, ruptures of the rotator cuff or biceps tendon, synovial chondromatosis (paper I-IV)</td>
</tr>
<tr>
<td>Fibromyalgia, rheumatic disease, major somatic or psychiatric disease (paper I-IV)</td>
</tr>
<tr>
<td>Patients that were not able to understand Norwegian (paper I-IV)</td>
</tr>
<tr>
<td>Patients that were unwilling to accept one of the treatment alternatives (paper III-IV).</td>
</tr>
</tbody>
</table>

The patients in the randomised controlled trial (paper IV) were recruited from April 2007 to December 2013. For the studies of agreement, reliability and validity of the different measurement tools (paper I and II), the inclusion period was from November 2006 to August 2008. Patients attending the responsiveness- and minimal clinical important change study (paper III), were included during November 2006 and March 2012.

3.2.2 Distribution of patients in the thesis

<table>
<thead>
<tr>
<th>Paper</th>
<th>Number of patients</th>
<th>Age</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>I and II</td>
<td>71</td>
<td>16-60</td>
<td>Isolated SLAP II lesions and recurrent anterior dislocations</td>
</tr>
<tr>
<td>III</td>
<td>89</td>
<td>18-60</td>
<td>Isolated SLAP II lesions</td>
</tr>
<tr>
<td>IV</td>
<td>120</td>
<td>18-60</td>
<td>Isolated SLAP II lesion</td>
</tr>
</tbody>
</table>

3.2.3 Randomisation and allocation to treatment (paper IV)

In the randomised controlled trial, all patients enrolled for treatment during arthroscopy were randomised into three groups (labral repair, mini-open biceps tenodesis, or placebo/diagnostic arthroscopy using the method of permuted blocks for random allocation, preformed by an independent statistician. This method was used to keep the numbers of subjects in the different groups in balance during the
inclusion period. Treatment allocation was preformed by an independent secretary who distributed sealed opaque numbered envelopes to the nurse manager in the operation theatre. The envelope was opened by a nurse only when the preoperative diagnostic evaluation had documented a type II SLAP lesion, and the inclusion criteria were met.

3.2.4 Clinical examination

The clinical examination before inclusion in the studies was performed by a manual therapist and/or an orthopaedic surgeon. At inclusion and during follow up, the clinical examination were performed by the manual therapist. Active and passive range of motion, pain within range of movement, scapulo-humeral movement pattern and strength were recorded. Clinical provocation testes for specific shoulder diagnostics and differential diagnosis were preformed, including O’Brien test / active compression test [48], the crank test [49], apprehension test [50], and Jobes relocation test [14] for possible SLAP lesions. Jerk test [51] was preformed to detect possible posterior labral lesions. Apprehension test and relocation test, load and shift test [50], sulcus sign [52], Gagey’s test [53] were used to classify the direction of any instability of the glenohumeral joint. Palpation of the long biceps tendon, Speed test [54], Yeargason test [55] and Bicep load test II [56], were preformed to detect pain from the long head of m.biceps brachialis. To detect pain from the acromioclavicular joint, palpation of the joint and the modified cross body test [57] were used. Possible tears or tendinopathies in the rotator cuff were searched by using isometric muscle tests, empty- and full can test (m.supraspinatus) [58], hornblowers sign (mm.infraspinatus and teres minor) [59] bear-hug test [60], belly-press test [61], and lift-off test [61] (m.subscapularis).

3.3 Procedures and measurements

To describe the patients in the studies and allow for comparison between studies, socio-demographic and background data consisting of gender, age, work status, physical activity level, which shoulder that is involved and dominant shoulder were described (paper I, II and III). In addition, type of treatment and capsular stiffness were reported (paper III).
Patients attending the reliability-, agreement- and validity-study (paper I and II) were examined by an experienced orthopaedic surgeon and an experienced manual therapist independently at the same day twice, with 7-10 days between the examinations, completing the 1988 version of the Rowe score. Other baseline registrations were obtained by the first examination by the manual therapist. The last examination was performed the day before start of the intervention (surgical treatment or active rehabilitation by physiotherapist). At the same time, patients completed the OISS, WOSI and EuroQol (EQ index and EQ-VAS).

In the responsiveness study (paper III), the patients were examined by an experienced manual therapist completing the 1988 version of the Rowe score and other baseline registrations the day before treatment and at 6 months follow up.

Patients attending the randomised controlled trial (paper IV) were examined by two experienced clinicians (one orthopaedic surgeon and one manual therapist) independently. The outcome measures and the baseline data (including gender, age, smoking, previous treatment, medication, physical activity level, work status, and presumed reason for the injury), were administrated by the manual therapist the day before surgery. Besides previously mentioned measures, the clinical Constant Murley score was used. This score is not mentioned in paper IV, but we decided to use it because the score is widely used for assessment in clinical trials regarding shoulder complaints and will give the opportunity for others to compare results. At 6 weeks follow up, the blinding of the patients is controlled by the manual therapist, who is blinded for the treatment, asking what treatment they perceive to have got. Range of movement, use of medicaments and postoperative stiffness are recorded by the manual therapist. At 3, 6, 24, and 48 months the patients will be assed by the blinded manual therapist completing the 1988 version of the Rowe Score and the Constant Murley score. Pain during activity and rest (over the last week) is recorded on a 0-100 visual analogue scale (VAS), comprising a horizontal ranging from no pain at one end to worst possible pain at the other end. The OISS, WOSI and the EuroQol will be completed by the patients at the follow up. Also the patients are asked to report change of main complaint on a scale from -9 (worst possible deterioration) to +9 (best possible improvement) [45].
3.3.1 Questionnaires and outcome measures

In paper I –III the outcomes evaluated are the OISS, WOSI, 1988 version of the Rowe score and EuroQol. The main outcomes in paper IV is the 1988 version of the Rowe score and the WOSI.

The 1988 version of the Rowe score is the latest of four versions of the score, and is an international recognized clinical score for evaluation of patients with shoulder instability [29-32;62]. The outcome measure consists of 5 domains; pain; stability; function; motion; and strength. The total score ranges from 0 (worst function) to 100 (best function). The outcome measure contains a patient evaluation (excellent, good, fair, and poor) for use at follow up.

The WOSI is a disease-specific, self-report questionnaire consisting of four domains; sports; recreation/work; lifestyle; and emotions) with total of 21 questions to be answered. Responses of each question are given on a visual analogue scale ranging from 0 (best function) to 100 (worst function). The total score ranges from 0 to 2100, where 2100 points indicates worst health related quality of life related to the shoulder [39].

The secondary outcomes in paper IV is the Constant Murley Score (CMS), the OISS, and the EuroQol (EQ-5D and EQ-VAS). The CMS is a combined patient and clinician reported outcome score for use in evaluation of patients with shoulder complaints. The score consists of 4 domains; pain; activities of daily living; active range of movement; and strength. The total score ranges from 0 (worst function) to 100 (best function) [28] .

The OISS is a self-report, disease-specific questionnaire [35]. The outcome measure consists of 12 questions with five response alternatives each. The response alternatives range from least to most difficulties (1-5 points). The items of the instrument cover episodes of instability, daily activities, pain, work, social life, sports/hobbies, attention to the shoulder problem, lifting, and lying positions, with a total possible score ranging from 12 to 60, where 60 points indicates worst health related quality of life related to the shoulder [35].

The EuroQol is a generic health-related quality of life outcome measure. The outcome consists of the EQ index also named EQ-5D, EQ-VAS and a register form for use in health-economic analysis. EQ-5D consists of 5 domains (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), with three levels
corresponding to no problem, some problem, and an extreme problem. The total score is ranging from the best imaginable state (1.0) to worst possible score (0.59). The EQ index is an estimation of generic health status using a visual analogue scale ranging from 0 (worst possible) to 100 (best possible) [34;63;64].

3.3.2 Translation
The EuroQol was already cross cultural adapted for use in a Norwegian-speaking population.[65]. The clinical Constant Murley score and the 1988 version of the clinical Rowe score were not translated and cross-cultural adapted since they were filled in by clinicians with sufficient understanding of English. Translation and cross-cultural adaptation of the OISS and WOSI was conducted according to recommended guidelines [66;67], including forward- and back translation by independent bilingual professional translators and medical researchers (paper I, II, and III).

3.3.3 Imaging
Patients suspected to have a type 2 SLAP lesions, had a magnetic resonance imagine with contrast liquid injected in the shoulder (MRI arthrography). MRI arthrography is reported to have a better specificity and sensitivity than MRI in diagnostics of SLAP lesions [68]. In addition, conventionally X-rays including outlet view and a clear projecting of the acromioclavicular joint were preformed to exclude patients with radiological and symptomatic arthrosis.

3.4 Statistics
3.4.1 Sample size calculation
For the observational studies evaluating agreement, reliability and validity (paper I and II) and the responsiveness and minimal clinical important change study (paper III), no formal sample size calculation was undertaken before starting the studies. A sample size of at least 50 subjects is recommended for a methods comparison study [69], and for estimation of minimal clinical important change [70]. In the randomized
The sham-controlled study (paper IV), the endpoints are at six and 24 months. The primary outcome measures are the 1988 version of the Rowe score and the WOSI. The minimal clinical important detectable difference for the 1988 version of Rowe score was not known at the planning of the study. This was not assessed for the WOSI for the actual diagnostic group. From clinical experience, we estimated that the minimal clinically important detectable difference is 10 points (0-100). Thus, we designed the trial to detect larger differences than 10 points between groups. The standard deviation was estimated to 15 points. To detect differences between treatment groups (SD = 15, \( \alpha = 0.05, \beta = 0.80 \), One-Way ANOVA) the study will require 36 patients in each group. Accounting for drop-outs we planned to include 40 patients in each group.

### 3.4.2 Measurement properties in the thesis

The thesis consists of an evaluation of several measurement properties for health-related outcome measures. Details are given in the table 3.4.2.1.
### 3.4.2.1 Table describing the measurement properties in the thesis

<table>
<thead>
<tr>
<th>Measurement properties</th>
<th>Definitions</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive analysis</td>
<td>Describes baseline values of the sample population</td>
<td>I-IV</td>
</tr>
<tr>
<td>Reliability:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-rater reliability</td>
<td>The degree of agreement between two or more raters [71]</td>
<td>I</td>
</tr>
<tr>
<td>Intra–rater reliability</td>
<td>The degree of agreement among multiple repetitions of a diagnostic test preformed by a single rater [71]</td>
<td>I</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>Measure of correlations among items measuring the same concept on questionnaire (sub)scales [70;71]</td>
<td>I and II</td>
</tr>
<tr>
<td>Measurement error</td>
<td>The systematic and random error of a patient’s score that is not attributed to true changes in the construct to be measured [71]</td>
<td>I and II</td>
</tr>
<tr>
<td>Agreement</td>
<td>Measure of absolute measurement error, addressing how close the scores of repeated measures are to another [70]</td>
<td>I and II</td>
</tr>
<tr>
<td>Validity:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content validity</td>
<td>Describes whether the concepts of interest are comprehensively represented by the items in the questionnaire [70;71]</td>
<td>I and II</td>
</tr>
<tr>
<td>Construct validity</td>
<td>Describes whether the questionnaire measures the relevant constructs related with other measures with theoretically derived hypotheses [41;70;71]</td>
<td>I and II</td>
</tr>
<tr>
<td>Convergent validity</td>
<td>Reflects correlations with other instruments that measures the same concepts, i.e. measures of constructs that theoretically should be related to each other are, in fact, observed to be related to each other [72]</td>
<td>I and II</td>
</tr>
<tr>
<td>Divergent/discriminant validity</td>
<td>Evaluation of whether concepts of measures that are supposed to be unrelated are in fact unrelated [72]</td>
<td>I and II</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>The ability to of an health related patient – reported outcome instrument (HR-PRO) instrument to detect changes over time in the construct to be measured [71]</td>
<td>III</td>
</tr>
</tbody>
</table>

### 3.4.3 Statistical methods

The specific statistical methods in this thesis, their definitions, descriptions and placement are described in the following tables.
### 3.4.3.1 Table describing the reliability statistical methods in the thesis

<table>
<thead>
<tr>
<th>Statistical methods</th>
<th>Definitions</th>
<th>Descriptions</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-retest reliability: Interclass correlation coefficient (ICC) version 2.1</td>
<td>Measurement of correlation between test and re-test combining 2-way random single measure. Express the ratio for within- and between patients variations</td>
<td>Values $&gt; 0.70$ are acceptable in comparing groups. Values $&gt; 0.90$ are acceptable in comparing individuals$^{[38]}$</td>
<td>I and II</td>
</tr>
<tr>
<td>Chronbach’s alpha</td>
<td>Measure of correlations among items measuring the same concept on questionnaire (sub)scales</td>
<td>A value between $0.70$ and $0.95$ indicates high correlation between items in a scale$^{[70;73]}$</td>
<td>I and II</td>
</tr>
</tbody>
</table>
### 3.4.3.2 Table describing the agreement statistical methods in the thesis

<table>
<thead>
<tr>
<th>Statistical methods</th>
<th>Definitions</th>
<th>Descriptions</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard error of the mean (SEM)/SE</td>
<td>An estimate of standard deviation of sample means; “a measure of the uncertainty of a single sample mean as an estimate of the population mean” [69]</td>
<td>Value with 95% CI given in the scale of the instrument</td>
<td>I and II</td>
</tr>
<tr>
<td>SEM&lt;sub&gt;agreement&lt;/sub&gt;</td>
<td>Extracting the square root of the residual mean square, using one-way analysis of variance with subjects as the factor [74-76] Takes the systematic differences between test and retest into account and equals the within-subject standard deviation (S&lt;sub&gt;w&lt;/sub&gt;) [75]</td>
<td>Value with 95% CI given in the scale of the instrument</td>
<td>I and II</td>
</tr>
<tr>
<td>Within-subject standard deviation (S&lt;sub&gt;w&lt;/sub&gt;)</td>
<td>Measurement of variation between test and retest within the same subject, and equals SEM&lt;sub&gt;agreement&lt;/sub&gt;</td>
<td>Value with 95% CI given in the scale of the instrument</td>
<td>I and II</td>
</tr>
<tr>
<td>Repeatability coefficient (RC)</td>
<td>Difference between two measurements/readings for the same subject, expected to be less than the repeatability coefficient of 95% of pairs of observations. Formula: RC = S&lt;sub&gt;w&lt;/sub&gt; x 1.96⁄√2. Equals minimal detectable change [70;73;77;78]</td>
<td>Value with 95% CI given in the scale of the instrument</td>
<td>I and II</td>
</tr>
<tr>
<td>Minimal detectable change (MDC)</td>
<td>Equals the repeatability coefficient.</td>
<td>Value with 95% CI given in the scale of the instrument</td>
<td>I –III</td>
</tr>
<tr>
<td>Limits of agreement (LoA)</td>
<td>Estimate of the difference of between the measurements for each subject (individual measurement error) with estimate of the interval within which 95% of differences would lie (95% confidence interval) [79]</td>
<td>Broad limits of agreement indicates high measurement error for individuals [79].</td>
<td>I and II</td>
</tr>
<tr>
<td>Limits of agreement plots</td>
<td>A graphical expression by a plot of difference against mean, exploring the relationship between difference and mean with 95% confidence interval [79].</td>
<td>Mean individual difference ± standard deviation (SD) of differences) using a logarithmic transformation. Diverging differences as the mean increase, indicates that the measurement error increases with the size of the measurement [79].</td>
<td>I and II</td>
</tr>
</tbody>
</table>
### 3.4.3.3 Table describing the validity statistical methods in the thesis

<table>
<thead>
<tr>
<th>Methods</th>
<th>Definitions</th>
<th>Descriptions</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floor and ceiling effects</td>
<td>The proportion of subjects reaching the lowest or highest possible score. [70:80]</td>
<td>Floor and ceiling effects were considered apparent in the studies if 15% or more of the responders had the lowest or highest possible score. Large floor and ceiling effects indicates that the content validity is low [70:80].</td>
<td>I and II</td>
</tr>
<tr>
<td>Hypotheses</td>
<td>Preliminary hypotheses testing construct validity, with expected mean differences between groups or expected correlations [70:80]</td>
<td>Mean differences or correlation measures; see below</td>
<td>I and II</td>
</tr>
<tr>
<td>Pearson’s correlation coefficient</td>
<td>Test of the strength of the linear association between the outcome and exposure variables</td>
<td>A correlation ((r) \geq 0.70) is considered as a positive correlation [70]</td>
<td>I and II</td>
</tr>
<tr>
<td>Student 2-sample t-test</td>
<td>Assessing differences between independent groups for continuous variables</td>
<td>Hypothesis rejected if (p&lt; 0.05)</td>
<td>I and II</td>
</tr>
<tr>
<td>Pearson’s (x^2) test</td>
<td>Assessing differences between groups for categorical variables</td>
<td>Hypothesis rejected if (p&lt; 0.05)</td>
<td>I and II</td>
</tr>
<tr>
<td>Independent sample t-test</td>
<td>Assessing differences between 2 populations</td>
<td>Hypothesis rejected if (p&lt; 0.05)</td>
<td>I and II</td>
</tr>
<tr>
<td>Spearman’s Rank correlation</td>
<td>Non-parametric correlation coefficient measuring the strength of association between two ranked variables</td>
<td>A correlation ((r) \geq 0.70) is considered as a positive correlation [70]</td>
<td>I and II</td>
</tr>
<tr>
<td>The multitrait-multimethod matrix</td>
<td>Testing divergent/discriminant validity</td>
<td>Formula: (r_{xy}/\sqrt{(r_{xx} \times r_{yy})}). A result &lt; 0.85 indicates acceptable discriminant validity [72]</td>
<td>I and II</td>
</tr>
</tbody>
</table>
3.4.3.4 Table describing the responsiveness statistical methods in the thesis

<table>
<thead>
<tr>
<th>Methods</th>
<th>Definitions</th>
<th>Descriptions</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized response mean (SRM)</td>
<td>$SRM_{\text{improved}} = \text{distribution-based responsiveness of the questionnaire.} \newline \newline SRM_{\text{unimproved}} = \text{the specificity to change}[75]$.</td>
<td>Dividing the mean change score (follow up minus baseline) by the standard deviation of the mean change. An effect size of $0.2 = \text{small, 0.5 = moderate, and 0.8 = large}[81]$.</td>
<td>III</td>
</tr>
<tr>
<td>Smallest detectable change (SDC)/ Minimal detectable change (MDC)</td>
<td>The specific value of the score of an instrument indicating a statistical improvement [82]</td>
<td>$SDC = SEM\text{agreement} \times 1.96 \times \sqrt{2}$, where 1.96 represents the z score 95% level of confidence [82].</td>
<td>I-III</td>
</tr>
<tr>
<td>Receiver operator curves (ROC)</td>
<td>Assessing of the sensitivity and specificity of correctly classifying patients as improved or unchanged according to an anchor.</td>
<td>A plot of true-positive rate (sensitivity) against the false-positive rate (1-specificity), showing the trade-off between the true-positive success and the false-positive error at each cut-off points in the change score.</td>
<td>III</td>
</tr>
<tr>
<td>Area under receiver operator curve (ROC$_{\text{AUC}}$)</td>
<td>The ability of the instrument to discriminate between subjects who are improved or unchanged according to an anchor.</td>
<td>A value of 1 = perfect accuracy. A value of 0.5 = by chance alone</td>
<td>III</td>
</tr>
<tr>
<td>Reliable change proportion (RCP)</td>
<td>Proportion of subjects improving by more than the smallest detectable change for each outcome measure [83].</td>
<td>Calculated by Wilson method. Differences between RCP with 95% confidence interval (CI) were calculated by Newcomb method for paired samples. 95% CI for RCP was obtained by the statistical program SAS</td>
<td>III</td>
</tr>
<tr>
<td>Minimal clinical important change (MIC)</td>
<td>The value of the score that illustrates a cut-off point for improved or unchanged according to an anchor</td>
<td>Integrated anchor-based and distribution based method. Patients were categorized as improved or unchanged according to a main complaint question (scale -9 to 9) and patient evaluation on Rowe score, where patients targeting “excellent” and “good” were categorized as improved and those targeting “fair” or “poor” where categorized as unchanged.</td>
<td>III</td>
</tr>
</tbody>
</table>
3.5 Ethical issues

In paper II and III, an ethical issue was the inclusion of patients under the age of 18 years. In this case, one of the parents of the patient signed the written consent. In paper I, no ethical issues were detected.

In the randomised controlled study (paper IV), the main ethical issue is the use of a placebo treatment (diagnostic arthroscopy) with postoperative physical treatment. The patients randomised into this group will have surgery with the risks including chance of allergic reaction for anaesthetics, infection and postoperative stiffness, without receiving a surgical treatment. All patients were informed of this possibility and had signed a written consent. With this placebo group there is a possibility that the patients would need a second surgical treatment in case physical treatment did not succeed in chancing their main complaint satisfactory. On the other hand, this possibility would be an option for patients receiving the two surgical treatments with postoperative physical treatment if they failed to get the patients satisfied with the result of the treatment. All the time there is no evidence favouring one of the three treatments in the study, the possibility of getting a satisfactory result should be equal.

Another ethical issue is the use of mini-open bicepstenodesis as a treatment. In this case, the labrum is not repaired, but the biceps tendon attached to the superior part of labrum and glenoid is transferred to the anterior part of the shoulder near by the bicipital groove and fixed with an anchor. The traction from the biceps tendon on the injured labrum is by this technique removed. This technique has been reported successful compared with labral repair [23], but these studies had no control group. A mini-open biceps tenodesis may cause postoperative complications as the other two treatments. Like the placebo group, patients receiving this treatment will not have a labral repair. The use of both diagnostic arthroscopy and mini-open biceps tenodesis with postoperative physical treatment can be justified since no studies have been able prove that arthroscopic repair is necessary for getting a satisfactory results [20].

The choice of six months as the first endpoint of the study and the first opportunity to break the blinding in case of non-satisfactory results may be a long time for patients not improving with the treatment they have been randomised to. The argumentation for this choice is that a new operation within six months may increase
the risk of having a infection in the shoulder, and that six months is considered to be a necessary time period to reach a satisfactory result, although this time period may be too short for getting a satisfactory result in some cases depending on postoperative stiffness and the patient perspective of a satisfactory result.
4 Synopsis

4.1 Paper I

Reliability, agreement and validity of the 1988 version of the Rowe Score
Skare Ø, Schröder CP, Mowinckel P, Reikerås O, Brox JI.

Background
The aim of this study was to evaluate reliability, agreement, and validity of the 1988 version of the Rowe score in patients with superior glenoid labrum lesions (SLAP) or recurrent anterior dislocations.

Methods
Thirty-eight patients with SLAP lesions and 33 with anterior recurrent dislocations were independently tested twice by 2 experienced clinicians with the 1988 version of the Rowe score. In addition, the patients completed disease specific and generic questionnaires the Western Ontario Shoulders Instability Index (WOSI), Oxford Instability Shoulder Score (OISS), and EuroQol (EQ-5D and EQ-VAS) at both test days.

Results
No significant differences between clinicians and test days were detected for the Rowe score.
Limits of agreement of the total score varied from -17.9 to 20.5. The within subjects standard deviation (equals SEMagreement) carried from 5.2 to 6.8. Standard error of mean was from 4.3 to 6.6. The minimal detectable change varied from 14.3 to 18.8. For the categorical score (total units), there were significant differences between the 2 examiners (P < .001). The ICC 2.1 was acceptable (>-.70) for the total score. No floor or ceiling effects were observed for the total score, but considerable floor and/or ceiling effects were detected for some of the domains. Discriminant or divergent
validity was acceptable, but content, construct, or convergent validity was not acceptable.

Conclusion
Results using the 1988 version of the Rowe score should be critically interpreted.

4.2 Paper II

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.

Skare Ø, Liavaag S, Reikerås O, Mowinckel P, Brox JI

Background
Having an estimate of the measurement error of self-report questionnaires is important both for assessing follow-up results after treatment and when planning intervention studies. Specific questionnaires have been evaluated for patients with shoulder instability, but not in particular for patients with SLAP (superior labral anterior posterior) lesions or recurrent dislocations. The aim of this study was to evaluate the agreement, reliability, and validity of two commonly questionnaires developed for patients with shoulder instability and a generic questionnaire in patients with SLAP lesions or recurrent anterior shoulder dislocations.

Methods
Seventy-one patients were included, 33 had recurrent anterior dislocations and 38 had a SLAP lesion. The patients filled in the questionnaires twice at the same time of the day (± 2 hours) with a one week interval between administrations. We tested the Oxford Instability Shoulder Score (OISS) (range 12 to 60), the Western Ontario Shoulder Instability Index (WOSI) (0 to 2100), and the EuroQol: EQ-5D (−0.5 to 1.0) and EQ-VAS (0 to 100). Hypotheses were defined to test validity.
Results
ICC ranged from 0.89 (95% CI 0.83 to 0.93) to 0.92 (0.87 to 0.95) for OISS, WOSI, and EQ-VAS and was 0.66 (0.50 to 0.77) for EQ-5D. The limits of agreement for the scores were: -7.8 to 8.4 for OISS; -339.9 to 344.8 for WOSI; -0.4 to 0.4 for EQ-5D; and -17.2 and 16.2 for EQ-VAS. SEMagreement for the total score were 2.9 for the OISS; 122.4 for the WOSI; 0.1 for the EQ-5D; and 6.0 for the EQ-VAS. The minimal detectable change for the total scores was 8.1 points for the OISS; 339.3 points for the WOSI; 0.4 points for the EQ-5D; and 16.6 points for the EQ-VAS. All questionnaires reflect the construct that was measured. The correlation between WOSI and OISS was 0.73 and ranged from 0.49 to 0.54 between the shoulder questionnaires and the generic questionnaires. The divergent validity was acceptable, convergent validity failed, and known group validity was acceptable only for OISS.

Conclusion
Measurement errors and limitations in validity should be considered when change scores of OISS and WOSI are interpreted in patients with SLAP lesions or recurrent shoulder dislocations. EQ-5D is not recommended as a single outcome.
4.3 Paper III

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

Skare Ø, Mowinckel P, Schröder CP, Liavaag S, Reikerås O, Brox JI
Shoulder & Elbow published online 27 May 2014. doi: 10.1177/1758573214534650

Background:
Evaluation of patients with SLAP lesions requires outcome measures validated for the purpose. The aim of the study was to evaluate 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 six months. Responsiveness was evaluated using standardised response mean (SRM), area under receiver operating characteristic curve (ROC_{AUC}), reliable chance proportion (RCP) statistics, and hypothesis. Minimal clinical important change (MCIC) estimates were reported.

Results:
All outcome measures had high values of SRM (0.86-1.92). RCP’s for the improved group were 68-79% for OISS, WOSI and Rowe score, and 15-49% for EuroQol. ROC_{AUC} 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.

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

STUDY PROTOCOL: Efficacy of labral repair, biceps tenodesis, and diagnostic arthroscopy for SLAP Lesions of the shoulder: a randomised controlled trial

Skare Ø, Schrøder CP, Reikerås O, Mowinckel P, Brox J
PMID:20929552

Background
Surgery for type II SLAP (superior labral anterior posterior) lesions of the shoulder is a promising but unproven treatment. The procedures include labral repair or biceps tenodesis. Retrospective cohort studies have suggested that the benefits of tenodesis include pain relief and improved function, and higher patient satisfaction, which was reported in a prospective non-randomised study. There have been no completed randomised controlled trials of surgery for type II SLAP lesions. The aims of this participant and observer blinded randomised placebo-controlled trial are to compare the short-term (6 months) and long-term (2 years) efficacy of labral repair, biceps tenodesis, and placebo (diagnostic arthroscopy) for alleviating pain and improving function for type II SLAP lesions.

Methods/Design
A double-blind randomised controlled trial are performed using 120 patients, aged 18 to 60 years, with a history for type II SLAP lesions and clinical signs suggesting type II SLAP lesion, which were documented by MR arthrography and arthroscopy. Exclusion criteria include patients who have previously undergone operations for SLAP lesions or recurrent shoulder dislocations, and ruptures of the rotator cuff or biceps tendon. Outcomes will be assessed at baseline, three, six, 12, and 24 months. Primary outcome measures will be the clinical Rowe Score (1988-version) and the Western Ontario Instability Index (WOSI) at six and 24 months. Secondary outcome measures will include the Shoulder Instability Questionnaire (SIQ), the generic EuroQol (EQ-5 D and EQ-VAS), return to work and previous sports activity, complications, and the number of re-operations.
Discussion

The results of this trial will be of international importance and the results will be translatable into clinical practice.

Trial Registration: [ClinicalTrials.gov NCT00586742]
5 General discussion

5.1 Patient sample and external validity

Patients attending the studies in this thesis were recruited from an orthopaedic outpatient clinic, and were referred to this clinic from physicians in general practice and other hospitals because they had shoulder complaints. The examinations of patients were performed in the outpatient clinic by orthopaedic surgeons or manual therapists. Patients with at history and clinical and radiographic signs of a SLAP lesion or recurrent anterior dislocations were invited to attend in the studies and assessed for inclusion by a manual therapist. There is a likely selection bias as the patients referred to the specialised shoulder clinic may differ from those who are treated in the primary care. Strict inclusion and exclusion criteria were applied for recruitment of patients with an isolated SLAP lesion or recurrent anterior dislocations. Patients with additional findings such as full thickness or symptomatic partial thickness rotator cuff tears, osteoarthritis of the glenohumeral joint, symptomatic osteoarthritis of the acromioclavicular joint, rheumatic disease affecting the shoulder, major somatic or psychiatric disease, and problems with understanding Norwegian language, were excluded. The main limitation affecting external validity is the degree of clinical transferable results as the inclusion criteria are narrow. An isolated SLAP lesion is less common than a SLAP together with rotator cuff pathology as observed in the present study. However, the aim of the randomised controlled study (paper IV) was to evaluated treatments for isolated SLAP lesions.

5.2 Patient sample and internal validity

The diagnosis of a SLAP lesion was obtained using clinical tests including the passive compression test/O’Brien test [48], the crank test [49], the apprehension test [31] with pain relief at the Jobe relocation test [84] and MRI arthrography. The accuracy of clinical tests used for diagnosing SLAP lesions is questionable, and single clinical shoulder tests are not recommended to diagnose a SLAP lesion [50]. Combination of clinical shoulder tests provides only marginal better accuracy, and the clinical diagnosis is made from history, clinical signs and imaging [50]. In the present studies MRI arthrography was used to confirm the diagnosis of a SLAP
lesion before arthroscopy. MRI arthrography has better sensitivity than MRI to detect SLAP lesions. The sensitivity (possibility of a test to be positive if there is a SLAP lesion) of MRI arthrography for diagnosing SLAP lesions has been reported to be between 73 and 100%. The specificity (the ability of the test to exclude a SLAP lesions when it is not apparent) have been reported to be between 69 and 98% [85]. The diagnosis was verified during arthroscopy because some patients have false positive MRI arthrography. These patients were excluded. The strict diagnostic criteria increase the internal validity of the studies (paper I-IV).

The diagnosis of recurrent anterior dislocations was obtained using patient history and defined as two or more episodes of dislocation. In addition, signs of clinical anterior instability was recorded using the apprehension test [31], the relocation test [84;86], and the load and shift test [87]. Other directions of instability were excluded with the jerk test [51], the sulcus sign [88] and the Gagey’s test [53] and the load and shift test [87]. Among the clinical tests suggesting anterior instability, the apprehension test has been reported to have the best likelihood ratio: 17 (CI 10.0 to 29.6) [50].

5.3 Study designs

The study designs chosen in this thesis differed depending on the purpose of each study included. In paper I and II, we used a test-retest design for assessment of the reliability and agreement. Patients in paper I was assed by two clinicians at the same time of the day within two hours and reassessed after seven days. An interval of seven days between test and re-test were chosen in order to reduce recall bias both in paper I and II. The order of assessment in paper I was identical, but the studies were not formally randomised, and this may have affected the results as the first assessment may influence the next. This effect would have been reduced by a proper randomisation between assessors. Patients were excluded if they reported considerable changes in shoulder condition between tests. Validity was evaluated using recommended methods [70;80].

A prospective cohort design was chosen in paper III. The study used a single group repeated measure design and patients were assessed before and after treatments [37]. Methodological problems in design of studies evaluating responsiveness are discussed [37;40;89;90]. The best design is to evaluate change
in measures within a randomised clinical trial, involving a treatment that is supposed to be efficacious. Alternatively, a single group measure design with evaluation of patients before and after a treatment is recommended [37]. Stratford et al claims that this is the weakest design in evaluation of responsiveness, and suggests the use of different hypothesis to support the estimates of important change [89]. In a review article form 2005, Stratford and Riddle [90] highlights the complexity of the use of multiple change coefficients applied on the same patient sample. They found that the detectable change for some questionnaires was smaller than the measurement error. They suggest the use of pilot studies to find the likely change characteristics of the population of interest as a guide for the choice of change coefficient. This strategy is vulnerable as the population of interest in a pilot study may be too small for such a purpose. The COSMIN group have recommended the use of prior hypotheses.

In paper IV, the protocol describes a randomised controlled double blinded design. This study design has been recommended as the best suitable design for comparing treatment effects [44;69]. The use of randomisation limits the allocation bias and systematic differences between groups [44;91]. Both known and unknown factors may influence the effect of a treatment. The randomisation process is the best method to ensure that the treatment given at the diagnostic arthroscopy is at random and not influenced by the surgeon’s beliefs. The block-randomisation ensures that the treatment given at the diagnostic arthroscopy is at random and not influenced by the surgeon’s beliefs. The block randomisation ensures that there are an approximately equal number of patients in each treatment group at any time in case the study has to be stopped earlier than planned.

Another main advantage of the design in paper IV is the inclusion of a sham control group. The purpose is that treatment expectations should be equal in the three interventions. All groups received postoperative standard physiotherapy to ensure that the only main difference between groups is the treatment given in the operation room. One limitation of the design in paper IV is the lack of a control group receiving no treatment. Ideally a no treatment group should have been included because the changes in the sham group may be attributed physiotherapy, the natural course, or placebo.

In paper IV, the blinding of the patient, the clinical assessor, and the physiotherapists giving postoperative treatment is also an important methodological advantage as it reduces bias that may contribute to treatment effects. The strategy
reduces the risk of influence of subjective expectations, which may influence the results [44]. By the same reason, the statistician and the researchers evaluating the study will be blinded for which treatment the three group of patients have received until the results are interpreted and consensus is reached.

The intention to treat analysis in paper IV means that the patients are evaluated according to the treatment group they were randomised into. If patients change treatment group during the trial, they will still be evaluated in their original treatment group. This strategy is critical to keep the effect of randomisation on comparable groups and to retain the power of the study which would drop if the patients were evaluated as treated.

Paper IV will give information about the effectiveness of the two types of surgical treatment and physiotherapy compared with physiotherapy alone. The study will however not give any information about physiotherapy compared to no treatment. A multi-centre study would have given the opportunity to include more patients, but is more difficult to administrate, and would imply more than one examiner for preoperative status and follow up, unless only self-reported questionnaires had been chosen, and also excluded the possibility of the same surgeon to perform all surgical treatments, which is considered to be a strength in the present study. However, the use of different surgeons and centres would have improved external validity.

5.4 Statistical approach

5.4.1 Sample size calculations

In the studies of agreement, reliability and validity (paper I and II) the sample size were larger than the minimum recommendations for such studies [69;70]. In the responsiveness study (paper III), there was no formal sample size calculation. The sample size was according to the recommendations of Altman for method comparison studies, 50 patients are assumed to be an adequate size [69]. Sample size calculations for such studies are difficult, especially when comparing different instruments, because the latter depends on which of the instruments the sample size calculations are based on. The obvious choice for sample size calculation in a randomised controlled trial is the primary outcome measure. In a method comparison study there is no primary end point, and the sample size calculation will vary
depending on the minimal detectable change. In a study of responsiveness, Stratford et al used a comparison of SRM between the scores [92]. Still, there is no agreement or clear recommendation in the literature of sample size in responsiveness studies [70;73]. We can not rule out that our sample size was too small since we did not provide a formal sample size calculation.

Sample size calculation in the randomised placebo controlled study (paper IV) was preformed using the 1988 version of the clinical Rowe score as one of two primary outcome measures. At the time of planning we did not know the minimal detectable change for the 1988 version of the Rowe score or the second primary outcome WOSI for the SLAP patients. Based on our experience from other studies [22;93], we decided that a clinical important change on the 1988 version of the Rowe score (0-100) is 10 points. We simulated multiple scenarios, and found that standard deviation (SD) between treatment groups was 14.6 units. To be able to detect a difference of 10 points between treatment groups with 95% probability and 80% power, the minimum study size was estimated to 36 patients per group. Assuming that there will be some possible dropouts, we planned to include 40 patients in each treatment group [94]. In the responsiveness study, the minimal clinical important change for the Rowe score was 17. Based on this finding, the randomised study should be well powered to detect a clinical significant difference.

5.4.2 Data collection

Baseline scores were usually recorded the day before the treatment started. The strength is that a change in the baseline score from inclusion to the start of the treatment might have influenced results. For the recording of test-retest scores (paper I and II), the time period of one week between tests diminish the effect of recall bias and reduce the possibility of change in the condition between tests. All patient-administrated scores were recorded in a paper version of the score. Use of an electronic recording device, not allowing patients to go to the next question before answering the prior, would have reduced the amount of missing of items.
5.4.3 Statistical methods

In paper I and II we reported both agreement and reliability as recommended [70;75]. As pointed out by de Vet et al, reliability is the most appropriate measure of reproducibility if the aim is the distinction of persons despite of measurement error, i.e. variability between study objects as in diagnostic use. To measure change in health status, agreement parameters are the most preferred, as small measurement error is required to distinguish clinical important changes from measurement error [70;75]. Reliability as measured by interclass correlation coefficient (ICC) is defined as between subjects variability divided on between subject’s variability and error, where error is the variability between time-points plus variability caused by random error. Different versions of ICC’s are often used to express reliability [75;95;96]. The weakness of ICC is that it will increase by reducing the variability of the sample (between patients variation), and by that reason be a misleadingly high value. This situation will occur if subjects differ little from each other (low between patients variation). A large difference within subjects can consequently be masked. In this thesis we used the ICC version 2.1, which is the ratio of variance derived from a two-way random model ANOVA. This method allow the error to be portioned between systematic and random error [96]. Agreement parameters may also be attributed to the same systematic error as reliability measures. In this thesis we used SEM agreement instead of SEM consistency to express within patient measurement error, because SEM consistency consists the ICC and SEM agreement (equals the within subject standard deviation) do not consists the ICC.

Also limits of agreement were used to express agreement since this method is not affected with the same problem as the use of ICC and SEM calculated from ICC.

In paper III we used different strategies to estimate responsiveness and minimal clinical important change. Although the COSMIN group has recommended to evaluate responsiveness using hypothesis method [73;80], there is no gold standard. The proposed guidelines from the COSMIN group have met disagreement [97]. In paper III, we chose to combine the some of the traditional distribution-based methods including SRM, area under ROC curve and RCP, with the anchor-based responsiveness and hypotheses focusing on change scores to reflect the different aspects of responsiveness. As limitation in this study was the inability to make
subgroups according to the state of changes; improved, unimproved, and deteriorated due to lack of power as discussed earlier.

In the placebo-controlled trial (paper IV), the intention-to-treat analysis is planned to maintain the power of the study. To avoid bias, patients, the evaluating clinicians, and the statistician are blinded for the treatment given. Any imbalance in baseline values will be adjusted to be able to compare the effect of treatment between groups using covariance. Besides the adjustment of baseline scores, this method is recommended in randomized trials because of the great statistical power [98].

5.4.4 Missing items

In the studies of agreement, reliability, validity and responsiveness of the outcome measures (paper I-III), patients with missing items were excluded in the analysis of the actual outcome measure since a sum score could not be obtained. There were no missing items in paper I and II. In paper III, eight patients were excluded because they did not complete the anchor. Two patients did not complete the WOSI, but were included for further analysis.

5.5 Outcome measures

Several types of outcome measures have been evaluated in this thesis. In the planning of the placebo-controlled trial, we discovered that no outcome measures had been developed or evaluated for use in patients with SLAP lesions. Our choice was to have both clinical outcomes and patient reported specific and general quality-of-life outcomes as we wanted to cover clinical measures as range of motion and strength together with the domains; pain; physical functioning; usual activities; emotional functioning; self-care; and patient ratings of improvement and satisfaction with treatment. The selection of outcomes was based on literature research. We decided to include numerical ratings of change in main symptoms and pain during rest and activity in order to perform a comprehensive evaluation of patients. The OISS and WOSI are measuring similar issues, despite of emotional functioning which is better covered in WOSI. The clinical measures of range of motion in the 1988 version of the Rowe were supposed to be beneficial in evaluation of patients with
restricted range of movement due to postoperative stiffness. In addition, we also included the Constant Murley Score in the placebo-controlled trial, since this measure is commonly used in shoulder research. This decision was made after the studies of psychometric properties of the outcome measures were started, and by this reason evaluation of this outcome measure is not included in these studies. Other outcome measures could have been included but would make the studies more difficult to administer.

5.6 Results

In paper I and II we evaluated reliability, agreement and validity of the 1988 version of the clinical Rowe Score, OISS, WOSI, and EuroQol in patients with SLAP lesions or recurrent anterior dislocations of the shoulder. We found that there was acceptable reliability for the total scores except for EQ-5D. The results for EQ-5D were in contrast with the findings of Adobor [99]. The results for EQ-VAS were in keeping with the result of Adobor et al [99]. For OISS, the results were in keeping with the results of Moser et al for OISS [100], though there was no information about which version of ICC that was used by Moser. ICC may vary depending on the version being used [96]. For WOSI, the test-retest reliability were in accordance with the original version and other language versions, with exception of the domain “sports, recreation and work” where Hatta et al [101] reported an ICC of 0.64 which is lower than we found (0.82, 95%CI 0.72 to 0.88). None of the other studies reported which version of ICC being used. There were acceptable reliability for the domains pain and stability in the Rowe score. In WOSI, all domains had acceptable reliability. The domains pain/discomfort and anxiety/depression were the only one that had acceptable reliability in EQ-5D. OISS and EQ–VAS do not consist domains.

Agreement statistics were acceptable and in the same range for the total scores of all outcome measures. The wide range in limits of agreement should be taken into account when interpreting the results for individuals. For Rowe score, the within subject standard deviation which equals $\text{SEM}_{\text{agreement}}$, was comparable with results reported for patients with rotator cuff disease using other outcomes [102], and with the other outcomes in this thesis. The limits of agreement of EQ-5D were between -0.4 to 0.4 in a scale ranging from -0.53 to 1, which indicates that this outcome is imprecise for estimating a true change in an individual patient. For OISS,
the results of measurement error were in keeping with the findings of Moser et al [100]. For the total scores of WOSI we found that measurement error expressed by $SE_{\text{agreement}}$ and MDC was is higher than reported in the Italian version [79]. The differences may attribute to versions of SEM, different diagnosis, and different methods to calculate MDC. There was no information about agreement in the original version or the other language versions [12;80-82]. Agreement statistics for the domains of the scores were not acceptable for the domain pain in Rowe score and the domains in EQ-5D.

The content validity of Rowe score may be questionable because the score was developed by an expert opinion not following modern criteria in development [70;80]. In addition the Rowe score include double-barrelled questions where patients are asked to answer more than one question at time, and considerable floor or/and ceiling effects for some of the domains in the score. The other outcomes reflect the construct to be measured. These outcomes are developed by modern criteria. There were no floor and ceiling effects for the total scores of OIIS, but floor and ceiling effects were observed for some of the single items. EQ-5D had considerable floor effects ranging from 22 to 97%. WOSI had no floor or ceiling effects.

The internal consistency for OIIS in was in keeping with results reported by of the developers(0.91) [35]. The results for the total scores of WOSI were similar or slightly better than reported by others [101;103-105]. For the domains of WOSI, we found stronger internal consistency for the domain lifestyle than reported by others [104;105]. This may be explained by the inclusion of patient with SLAP lesions in the present study or the size of the study. Patients with SLAP lesions may differ from those with anterior instability regarding this domain of lifestyle. The study populations were limited in the previous versions and counted 22 patients in the Swedish version [105] and 25 in the German version [104], while the present study [106] counted 71 patients. The internal consistency for EQ-VAS was in agreement with results reported by others, but slightly lower for EQ-5D [99].

Validity evaluated by hypotheses was acceptable only in terms of discriminant/divergent validity for all outcomes. Convergent validity failed for all outcomes, but known group validity was acceptable for OIIS. No other comparable studies had used hypothesis for evaluation of validity. To ease comparison with other studies using correlations as measure of construct validity, we compared all the outcomes. OIIS and WOSI had acceptable correlations 0.64 (0.41 to 0.80) for the
SLAP group, and 0.80 (0.62 to 0.69) for the instability group. No other correlations were above 0.60. Although, such comparisons are frequently used for evaluation of construct validity, the value of the method is questionable [71;73;80].

Responsiveness was evaluated by distribution-based methods, anchor based methods, and hypothesis method. We found that the self-reported OISS, WOSI, and the clinician assessed Rowe score are responsive in evaluation of patients with SLAP lesions. The OISS and Rowe score seems to be the most sensitive measures. The anchor-based results suggested that the generic EuroQol is less responsive. The anchors discriminated well between those who considered themselves improved compared with those who did not. The distribution-based methods using SRM and RCP indicated relatively small differences in responsiveness between the clinical assessed Rowe score and self reported shoulder specific questionnaires OISS and WOSI. The values of SRM for EQ-5D and EQ-VAS was acceptable. Contrary, the anchor based methods MCIC and ROC analysis and the hypotheses indicated that the Rowe score and the shoulder specific questionnaires are more responsive than the generic EuroQol. The distribution-based method may overestimate the responsiveness of EQ-5D while the anchor-based methods suggest 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. These findings highlight the possible differences in results depending on which method that are used for evaluation of responsiveness. The Rowe score was superior compare with the other outcomes in differencing between patients with or without capsular stiffness. The MCIC estimates may be difficult to distinguish from measurement error for the Rowe score. For OISS, WOSI and the EuroQol, the MCIC were considerable lower than the measurement error. MCIC values and sensitivity measures, suggests that OISS and Rowe score are superior compare with the other outcomes. Comparison of the results of responsiveness with other studies is difficult because of very little information about responsiveness in the other studies. For OISS Moser et al reported ROC$_{AUC}$ of 0.80 which is a little lower than our findings of 0.92. The SRM values for WOSI of 1.73 is higher than reported by Salomonsen et al (SRM 1.40)[105], Oh et al (0.66) [46], and the developers (0.93) [39].

Paper III contributes to the knowledge of responsiveness of the tested outcome measures, but is important to remember that responsiveness and especially
MCIC is highly depending on many factors in a study, and the value of comparing results between studies may be limited.

The results of the sham controlled trial (paper IV) are still not analysed since the follow up is not completed. In the planning of this study, we were concerned about the use of a sham-group. We feared that patients would hesitate to join the study because of the risk of receiving no surgical treatment. During the inclusion process, we were surprised to find that very few patients refused to join the study because of the risk of being included in the sham-group. The main challenges were to find patient with isolated SLAP II lesions and the resistance in other orthopaedic clinics to send patients with SLAP lesions for possible inclusion. Another concern in the planning of the study was the ethical aspect of giving a sham treatment with possible risks as postoperative stiffness and infection. However, sham treatment has been documented as beneficial for patients [107;108]. Taken this evidence into account, the use of a placebo is not more unethical than to recommend the patients operations that are not evidence based in terms of evaluated by randomised controlled trials.
6 General conclusions

**Paper I**

The aim of the study was to evaluate reliability, agreement and validity for the clinical Rowe score on patients with SLAP lesions and recurrent dislocations using recommended statistical methods. The results suggest that the 1988 version of the Rowe score is reliable for evaluation of patients with SLAP lesions and recurrent dislocations. The measurement error (agreement) may indicate that the score is better suited for evaluation of patients in a group than on an individual level. Divergent validity was satisfactory, while content validity and convergent validity failed.

**Paper II**

The aim of the study was to evaluate agreement, reliability and validity of two commonly used questionnaires developed for patients with shoulder instability, and a generic questionnaire in patients with type II SLAP lesions or recurrent anterior dislocations. Measurement error on an individual level was considerable, but acceptable on a group level, indicating that the score is better suited for evaluation on a group level than in an individual level. Reliability was acceptable for all outcome measures except for EQ-5D. Content validity and divergent validity were satisfactory. Convergent validity failed.

**Paper III**

The aim of the study was to evaluate responsiveness and minimal clinical important change of the 1988 version of the Rowe Score, Oxford Instability Shoulder Score (OISS) and Western Ontario instability Index (WOSI), and EuroQol in patients with type 2 SLAP lesions. Distribution-based methods for evaluation of responsiveness suggested that all outcome measures are responsive, but anchor-based methods and hypotheses indicates that the Rowe score and the OISS and WOSI are more responsive than the EuroQol in patients with type 2 SLAP lesions.
General conclusion

The 1988 version of the Rowe Score and the Norwegian translated versions of OISS and WOSI have adequate reliability and responsiveness for evaluation of patients with SLAP lesions. The relatively large measurement errors suggest that these outcome measures are better suited for evaluation of patients at a group level than at an individual level. This should be taken into account when individual patients are informed about the expected effect of the treatment. The divergent validity of the outcomes was acceptable suggesting that constructs with different meanings can be discriminated from each other. Convergent validity failed, which means that similar concepts are difficult to compare between scores. Though there were minor differences between the patient reported OISS and WOSI and the clinical Rowe score in psychometric properties, the OISS is easier to administer and can be recommended as a single outcome measure. The EQ-5D is not recommended as a single outcome measure because of the limited reliability, considerable measurement error, limited validity and responsiveness. The psychometric properties of EQ-5D should be considered when using this outcome measure in cost benefit analysis. For use in the randomised, double blinded, sham–controlled trial (paper IV), the primary outcomes Rowe score and WOSI and the secondary outcome OISS seems to be suited for evaluation of patients with type 2 SLAP lesions.
References

Reference List


(10) OLSSON O. Degenerative changes of the shoulder joint and their connection with shoulder pain; a morphological and clinical investigation with special attention to the cuff and biceps tendon. Acta Chir Scand Suppl 1953;181:1-130.


Ref Type: Serial (Book,Monograph)


(64) Essink-Bot ML, Krabbe PF, Bonsel GJ, 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 May;35(5):522-37.


Ref Type: Serial (Book,Monograph)


(71) Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and

(72) CAMPBELL DT, FISKE DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 1959 Mar;56(2):81-105.


Ref Type: Serial (Book,Monograph)


(99) Adobor RD, Rimeslatten S, Keller A, Brox JI. Repeatability, reliability, and concurrent validity of the scoliosis research society-22 questionnaire and EuroQol in


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

Øystein Skare1*, Sigrud Liavaag2, Olav Reikerås3, Petter Mowinckel4 and Jens Ivar Brox3

Abstract

Background: Having an estimate of the measurement error of self-report questionnaires is important both for assessing follow-up results after treatment and when planning intervention studies. Specific questionnaires have been evaluated for patients with shoulder instability, but not in particular for patients with SLAP (superior labral anterior posterior) lesions or recurrent dislocations. The aim of this study was to evaluate the agreement, reliability, and validity of two commonly questionnaires developed for patients with shoulder instability and a generic questionnaire in patients with SLAP lesions or recurrent anterior shoulder dislocations.

Methods: Seventy-one patients were included, 33 had recurrent anterior dislocations and 38 had a SLAP lesion. The patients filled in the questionnaires twice at the same time of the day (± 2 hours) with a one week interval between administrations. We tested the Oxford Instability Shoulder Score (OISS) (range 12 to 60), the Western Ontario Shoulder Instability Index (WOSI) (0 to 2100), and the EuroQol: EQ-5D (−0.5 to 1.0) and EQ-VAS (0 to 100).

Hypotheses were defined to test validity.

Results: ICC ranged from 0.89 (95% CI 0.83 to 0.93) to 0.92 (0.87 to 0.95) for OISS, WOSI, and EQ-VAS and was 0.66 (0.50 to 0.77) for EQ-5D. The limits of agreement for the scores were: -7.8 to 8.4 for OISS; -339.9 to 344.8 for WOSI; -0.4 to 0.4 for EQ-5D; and −17.2 and 16.2 for EQ-VAS. All questionnaires reflect the construct that was measured. The correlation between WOSI and OISS was 0.73 and ranged from 0.49 to 0.54 between the shoulder questionnaires and the generic questionnaires. The divergent validity was acceptable, convergent validity failed, and known group validity was acceptable only for OISS.

Conclusion: Measurement errors and limitations in validity should be considered when change scores of OISS and WOSI are interpreted in patients with SLAP lesions or recurrent shoulder dislocations. EQ-5D is not recommended as a single outcome.

Keywords: Outcome measurements, Reliability, Agreement, Validity, Oxford instability shoulder score, Western ontario shoulder instability index, EuroQol, SLAP-lesions, Shoulder dislocations
Background

A number of self-report questionnaires have been developed to assess shoulder pain and disability from the patient’s perspective. The choice of a questionnaire may be based on factors such as study or diagnostic group, practical considerations regarding the ease of scoring, and the time to fill in the questionnaire as well as clinometric properties. A recent study reported that a general shoulder questionnaire was as good as the disease specific Western Ontario Rotator Cuff Index (WORC) and Oxford Shoulder Score (OSS) for rotator cuff disease [1]. Thus, the need for disease specific questionnaires for all different kinds of shoulder diagnoses can be questioned.

Shoulder instability can be defined as the loss of shoulder comfort and function due to undesirable translation of the humeral head on the glenoid [2]. From the patient perspective, shoulder instability may be defined as symptomatic abnormal motion of the glenohumeral joint which can present as pain or a sense of displacement (subluxation or dislocation) [3]. From a diagnostic point of view, instability is difficult to verify unless a dislocation has occurred. The latter is defined as a complete dissociation of the articular surfaces documented radiographically or by a manual reduction manoeuvre [4]. In patients with anterior shoulder dislocation, the main patho-anatomical finding is the Bankart lesions with avulsion of the labrum and the glenohumeral ligament from the anterior-inferior glenoid rim. A superior labral anterior posterior (SLAP) lesion of the shoulder is a relatively rare condition caused by injury or degeneration of the superior part of the glenoid labrum. Apprehension and loss of confidence are reported to be the major factors inhibiting sports activities and decreasing quality of life in patients with recurrent dislocations [5,6], while pain, popping, clicking, catching, weakness, stiffness, and instability (apprehension and loss of confidence) are reported in patients with SLAP lesions [7].

Several questionnaires have been designed to evaluate treatment of instability in the shoulder while specific questionnaires have not been published for patients with SLAP lesions. In the original study the Western Ontario Shoulder Instability Index (WOSI) was evaluated in 33 patients with shoulder instability, but not in particular for patients with recurrent shoulder dislocations [2]. Oxford Instability Shoulder Score (OISS) was evaluated in 53 patients diagnosed as having either unidirectional or multidirectional instability [6]. In a 5-year follow-up study of arthroscopic repair in patients with SLAP lesions [8], the clinical Rowe Score (1988 version) was used as the main effect variable. This score has been reported to have considerable limitations [9] and results [8] would have been strengthened applying a self-report outcome with acceptable measurement properties.

In absence of a disease-specific scoring system for SLAP lesions, existing questionnaires for shoulder instability [10], such as the OISS [11] and the WOSI [2], offer a possible alternative for the assessment of treatment effects in patients with SLAP lesions, because both conditions includes labral lesions that may cause similar symptoms.

The purpose of the present study was to cross-culturally adapt OISS and WOSI for use in Norwegian-speaking patients, and evaluate the agreement, inter-rater and intra-rater reliability, content- and construct validity of the Oxford Instability Shoulder Score, the Western Ontario Shoulder Instability Index, and the EuroQol in patients with recurrent anterior shoulder dislocations or SLAP lesions.
Methods

Study population and study design
Between November 2006 and August 2008, 103 patients referred for shoulder surgery at the Orthopaedic Department at Lovisenberg Diocesan Hospital in Oslo, Norway, were prospectively recruited. Eighty-five patients aged 16–60 years with a symptom duration of at least 3 months met the inclusion criteria for the study [9]. All patients signed an informed consent. The present study is approved by The Ethical Committee of Health Region South-East, Norway. Seventy-one patients (33 had recurrent anterior (at least two) dislocations and 38 had a SLAP lesion) were included. Patients with symptoms and signs suggesting a SLAP lesion were included if the lesion was confirmed on MRI arthrography [9]. Patients labelled SLAP lesion were not included if they had a history of shoulder dislocation. The exclusion criteria for the study were posterior or multidirectional dislocations; inability to complete the questionnaires; previous surgery for SLAP injuries or instability in the same shoulder; rheumatic disease affecting the symptomatic shoulder; pain referred from the cervical or thoracic spine; and severe somatic or psychiatric disorders. All included patients gave a written informed consent.

The patients completed OISS, WOSI, the 1988 version of Rowe Score, and EuroQol questionnaire twice, at the same time of the day with a one week interval between administrations. The test-retest period was chosen to reduce recall bias. One patient was excluded at retesting because he reported major changes in his activity level, and deterioration between tests.

Questionnaires
OISS is a disease-specific health-related quality-of-life self-report questionnaire, for use in patients with shoulder instability [6]. Several names and abbreviations have been used synonymously, such as Oxford Instability Score (OIS) [19] and Shoulder Instability Questionnaire (SIQ) [20]. The instrument consists of 12 questions, each of which had five response alternatives, ranked from least to most difficult (1 point). The items cover episodes of instability, daily activities, pain, work, social life, sports/hobbies, attention to the shoulder problem, lifting, and lying positions with a total possible score ranging from 12 (best function) to 60 (worst function) [6].

WOSI consists of 21 self-report questions representing four domains (sports, recreation/work, lifestyle and emotions). Each question is answered on visual analogue scale ranging from 0 (best) to 100 (worst). The total score ranges from 0 (best) to 2100 (worst) [2].

The EuroQol is a generic health-related quality-of-life instrument [12,21,22]. EQ-5D consists of five domains (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), with three levels corresponding to no problem, some problem and an extreme problem. The responses are transformed into a utility index and are then classified into 243 (35) health states ranging from the best imaginable state (1.0), and worst possible score (0.59). EQ-VAS estimates generic health status by using a visual analogue scale from 0 (worst possible) to 100 (best possible).

For assessment of the correlation between scores we also included the 1988 version of the clinical Rowe Score [9,23].

Translation
The EQ-5D was already cross-culturally adapted for use in Norwegian-speaking population [24]. Cross-cultural adaptations of the Norwegian versions of OISS and WOSI was conducted according to the procedures described in the literature [25,26]. Forward translation of OISS and WOSI was done by two bilingual medical doctors, one bilingual nurse and one bilingual medical doctor and professional translator. Two had Norwegian as their native language and two had English as their native language. The translations were done independent of each other and then compared. The Norwegian versions were then back-translated into English by a professional translator. The back-translated versions were then reviewed and inconsistencies of the items of OISS and WOSI were discussed and approved in a consensus meeting with the four translators.

Statistical analysis
The study was planned to have a sample size of at least 50 patients, which is the general recommendation given by Altman for a methods comparison study [27]. All patients had chronic complaints and we assumed that diagnostic group did not influence agreement statistics. For reliability and validity evaluation we could not exclude that diagnostic group may influence results and some exploratory analyses were performed in each diagnostic group.

Age, duration of symptoms, and number of dislocations were described by median (range) while numbers (percentages) are reported for gender, manual labour, physical activity level, and whether the dominant shoulder was involved. Means (SD) were used for descriptive statistics for total scores and domain scores of WOSI and for the total scores of OISS, EQ-5D and EQ-VAS.

The data of the descriptive statistics data followed a normal distribution. Differences between groups were compared by Student’s two-sample t-test, Chi-square was used for categorical variables. Minimum and maximum scores for individual items, domain and total scores were examined for possible floor and ceiling effects, which were considered to be present if more than
15% of respondents achieved the highest or lowest score, respectively.

Internal consistency describes the correlations among items measuring the same concept on questionnaire (sub)scales [17], A Chronbach's alpha between 0.70 and 0.95, indicates strong correlation between items in a scale [17,18]. We calculated the internal consistency for the total scores and domain scores.

Test–retest reliability is commonly tested by ICC, which combines the within and between patient variation from 0 (no reliability) to 1 (perfect reliability). According to Terwee et al., an ICC > 0.70 is considered to be acceptable [17]. We used a two-way random single measure (ICC 2.1), with a 95% confidence interval for the total score and for the domains [17,28].

Agreement describes the within patient measurement error, and indicates how close the scores of repeated measurements are to one another [17]. Statistical methods to estimate measurement error include standard error of measurement (SEM), limits of agreement (LoA), and minimal detectable change (MDC) which equals the repeatability coefficient [17,18,29]. SEM is recommended as the measure of agreement [18]. It can be estimated as SEM\textsubscript{consistency} (SD√(1-ICC)) or SEM\textsubscript{agreement} (within- subject standard deviation (S\textsubscript{w})). The latter is obtained by extracting the square root of the residual mean square, using one-way ANOVA with subjects as the factor [30,31]. While the SEM\textsubscript{consistency} include both between and within-subject variations, SEM\textsubscript{agreement} takes only the within-subjects variation into account. The COSMIN checklist for does not give information about a particular version of SEM [18,32,33]. In the present study, we estimated SEM\textsubscript{agreement} minimal detectable change (SEM × 1.96√2) and limits of agreement (mean individual difference ± SD of differences) with 95% confidence interval. We constructed agreement plots according to Bland and Altman [34].

Validity describes whether an instrument measures what it is intended to [13].

Content validity indicates that the concepts of interest are comprehensively represented by the items in the questionnaire [32,35]. Terwee et al. recommended that authors should provide clear descriptions aims of the questionnaire, the target population, the concepts intended to be measured, item selection, reduction and interpretability [17]. According to the COSMIN checklist [32], content validity should be assessed by making a judgment about the relevance and comprehensiveness of the items. Patients or experts should be asked whether they missed any items. In the present study, this was checked during the cross cultural adaptation process and by assessing floor and ceiling effects of the domains and single questions of the instruments [17,32]. Large floor and ceiling effects suggest that content validity is low. Floor and ceiling effects were considered apparent if 15% or more of the responders had the lowest or the highest possible score, respectively.

Construct validity means that questionnaire measures the relevant constructs [33]. The COSMIN checklist recommends to use hypotheses to test relationships with other instruments or differences among relevant groups [32]. Construct validity is considered acceptable when at least 75% of the hypotheses are accepted [17]. To admit comparison of construct validity with other studies not using hypotheses, Pearsons correlation coefficient between OISS, WOSI, EQ-5D, EQ-VAS and the 1988 version of Rowe Score was obtained.

There are several aspects of construct validity which include convergent, divergent/discriminant, and known group validity. Convergent validity reflects correlation with other instruments that measure the same properties [39]. Convergent validity for hypotheses 1 to 8 was tested using Pearsons correlation coefficient. R > 0.70 was regarded as positive correlation [17]. Divergent validity/discriminant validity evaluates whether concepts of measures that are supposed to be unrelated are in fact unrelated [36]. Tests can be invalidated by too high correlations with other tests they were intended to differ [36]. In the present study the formula \(r_{xy} / √(r_{xx} * r_{yy})\) was used to test discriminant validity [36]. Hypotheses 12 and 13 were tested using the formula \(r_{xy} / √(r_{xx} * r_{yy})\), where \(r_{xy}\) is the correlation between EQ-5D and OISS and WOSI, \(r_{xx}\) is the ICC of OISS or WOSI, and the \(r_{yy}\) is the ICC of EQ-5D. A result <0.85 is considered to indicate acceptable discriminant validity [36]. Known group validity describes the relationships among different groups (age, gender, diagnosis, etc.). Independent sample t- tests were used to test known group validity for hypotheses 9 to 11.

Hypotheses

Convergent validity (positively correlated means \(r > 0.70\))

1. WOSI should be positively correlated OISS.
2. WOSI should be positively correlated with Rowe Score.
3. OISS should be positively correlated with Rowe Score.
4. WOSI part B (Sports/Recreation/Work) should be positively correlated with question 8 of OISS:
   “During the last four weeks, how much has the problem with your shoulder interfered with your sporting activities or hobbies?”
5. WOSI part D (Emotions) should be positively correlated with question 9 of OISS: “During the last four weeks, how often has your shoulder been «on your mind» - how often have you thought about it?”
6. WOSI part C (Lifestyle) should be positively correlated with question 12 of OISS: “During the last four weeks, have you avoided lying in certain positions in the bed at night because of your shoulder?”

7. WOSI part A (Physical symptoms) should be positively correlated with question 3 of OISS: “During the last three months, how would you describe the worst pain you have had from your shoulder?”

8. Question 1 of OISS — “During the last six months, how many times has your shoulder slipped out of joint (or dislocated)?” — should be correlated with question 8 of WOSI part A: “How much feeling of instability or looseness do you experience in your shoulder?”

**Known group validity**

9. OISS should be the same for patients < 45 and > 45 years old.

10. WOSI should be the same for patients < 45 and > 45 years old.

11. The scores of the SLAP group should be negatively correlated (R < 0.70) with the scores of the instability group of question 1 of OISS: “During the last six months, how many times has your shoulder slipped out of joint (or dislocated)?”

**Divergent/discriminant validity**

12. The discriminate validity between OISS and EQ-5D should be < 0.85.

13. The discriminate validity between WOSI and EQ-5D should be < 0.85. The analysis was performed using Statistical Analysis System software (SAS, version 9.2, SAS Institute Inc., Cary NC, USA).

**Results**

**Demographics**

Fifty men (70.4%) and 21 women (29.6%) were included for further analysis in this study (Table 1). There were no differences in baseline characteristics among the 14 patients who were excluded, compared with those patients who were included. The patients in the instability group were younger than the SLAP group and had a median of 10 (range 2 to 40) dislocations. The two diagnostic groups did not differ on the mean scores of the questionnaires.

**Cross cultural validity**

The EuroQol instrument was already cross-culturally adapted into Norwegian [24]. The relevance and translations of items of OISS and WOSI were discussed and approved by the consensus group. The translated versions of OISS and WOSI adequately reflected items in the original-language versions.

**Internal consistency**

Chronbach’s alpha for the total scores of OISS, WOSI, and EQ-VAS was ranged from 0.94 to 0.96 (Table 2). There Chronbach’s alpha was 0.79 for EQ index and ranged from 0.87 to 0.96 for the domains of WOSI (Table 2).

**Test-retest reliability**

ICC ranged from 0.89 (95% CI 0.83 to 0.93) to 0.92 (0.87 to 0.95) for the total scores of OISS, WOSI, and EQ-VAS and was 0.66 (0.50 to 0.77) for EQ-5D (Table 2). For the domains of WOSI, ICCs ranged from 0.77 (0.65 to 0.85) to 0.92 (0.88 to 0.95) (Table 3). ICC ranged from 0.01 (−0.22 to 0.24) to 0.75 (0.63 to 0.84) for the domains: walking, personal care, and daily activities of EQ-5D (Table 4).

**Agreement**

There were no significant differences between the first and second administration of the scores (Table 2) or between diagnostic groups (Table 1). SEMagreement for the total score were 2.9 for the OISS; 122.4 for the WOSI; 0.1 for the EQ-5D; and 6.0 for the EQ-VAS (Table 2). The minimal detectable change for the total scores was 8.1 points for the OISS; 339.3 points for the WOSI; 0.4 points for the EQ-5D; and 16.6 points for the EQ-VAS (Table 2). For the total scores, the limits of agreement were −7.8 to 8.4 for the OISS; -333.9 to 344.8 for the WOSI; -0.4 to 0.4 for the EQ-5D; and −17.1 to 16.2 for the EQ-VAS (Table 2). For the domains of the WOSI and the EQ-5D, the results are given in Tables 3 and 4 respectively. The limits of agreement plots are shown in Figure 1.

**Content validity**

The OISS, the WOSI, and the EuroQol reflected the construct to be measured. However, in this study, 4 of the 38 patients with SLAP lesions reported experiencing shoulder dislocation over the previous 6 months (Question 1, OISS). There were no floor and ceiling effects for the total score of OISS or the single item scores, the domain scores, and the total score of WOSI. For single items of OISS, floor effects were observed for question 1 (shoulder instability) in the SLAP group and in both groups for question 2, 7, and 12, and ceiling effects for question 7, 9, 10, and 12. For EQ-5D the floor effects ranged from 22% to 97% (Table 4).
**Table 1 Descriptive statistics**

<table>
<thead>
<tr>
<th></th>
<th>SLAP</th>
<th>Instability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males/females [n]</td>
<td>28/10</td>
<td>22/11</td>
</tr>
<tr>
<td>Age (median [range])</td>
<td>40 (16–60)</td>
<td>25 (19–54)</td>
</tr>
<tr>
<td>Duration of symptoms median months (range)</td>
<td>23 (4–132)</td>
<td>36 (10–360)</td>
</tr>
<tr>
<td>Manual labour n (%)</td>
<td>21 (55.3)</td>
<td>14 (42.4)</td>
</tr>
<tr>
<td>Physical activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>competition</td>
<td>4 (10.5)</td>
<td>5 (15.2)</td>
</tr>
<tr>
<td>weekly or more</td>
<td>20 (52.6)</td>
<td>20 (60.6)</td>
</tr>
<tr>
<td>none</td>
<td>14 (36.8)</td>
<td>8 (24.2)</td>
</tr>
<tr>
<td>Shoulder involved; right/left</td>
<td>27/11</td>
<td>14/19</td>
</tr>
<tr>
<td>Dominant shoulder involved n (%)</td>
<td>26 (68)</td>
<td>15 (45)</td>
</tr>
<tr>
<td>Number of dislocations median (range)</td>
<td>0</td>
<td>10 (2–40)</td>
</tr>
<tr>
<td>WOSI total score</td>
<td>1081.7 (382.8)</td>
<td>1025.8 (438.9)</td>
</tr>
<tr>
<td>OISS total score</td>
<td>37.4 (7.6)</td>
<td>33.7 (10.4)</td>
</tr>
<tr>
<td>EQ-SD index</td>
<td>0.65 (0.22)</td>
<td>0.76 (0.25)</td>
</tr>
<tr>
<td>EQ-VAS</td>
<td>71.2 (15.0)</td>
<td>72.7 (21.3)</td>
</tr>
<tr>
<td>Rowe total score</td>
<td>66.9 (10.6)</td>
<td>63.9 (11.0)</td>
</tr>
</tbody>
</table>

SLAP superior glenoid labrum lesions, WOSI Western Ontario shoulder Instability Index, OISS Oxford Instability Shoulder Score, EQ-SD, EQ-VAS, EuroQol. Rowe score; 1988 version. Scores are given for first evaluation.

**Construct validity**

There were no missing items. The correlation between WOSI and OISS was: 0.64 (95% CI 0.41 to 0.80) for the SLAP group and 0.80 (95% CI 0.62 to 0.69) for recurrent dislocations. The correlations between the specific questionnaires and EQ-SD and EQ-VAS ranged from −0.27 (95% CI −0.54 to 0.05) to −0.59 (95% CI −0.79 to 0.32) with r < 0.60 for both diagnostic groups. The Rowe score correlated −0.42 (95% CI −0.67 to −0.09) with WOSI for the SLAP group, and −0.59 (95% CI −0.76 to −0.33) for the recurrent dislocation group, r < 0.60 in both groups. The correlation between the Rowe score and OISS was −0.30 (95% CI −0.58 to 0.05) for the recurrent dislocation group, and −0.45 (95% CI −0.67 to −0.15) for the SLAP group r < 0.60 in both groups.

**Convergent validity**

Hypotheses (1 to 8) failed (r > 0.70 only for hypothesis 1).

**Known group validity**

Hypotheses (9 to 11) failed (p<0.05 only for OISS, hypothesis 9).

**Divergent/discriminant validity**

Hypotheses 12 and 13 were accepted, with r = 0.58 and 0.57, respectively.

**Table 2 Agreement and reliability statistics - total scores**

<table>
<thead>
<tr>
<th></th>
<th>1.test Mean (SD)</th>
<th>2.test Mean (SD)</th>
<th>Mean difference (95% CI)</th>
<th>Limits of agreement (LoA)</th>
<th>Minimal detectable change (95% CI)</th>
<th>ICC (2.1) (95% CI)</th>
<th>Standard error of measurement (SEM)</th>
<th>Chronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>OISS (12 to 60)</td>
<td>35.7 (9.1)</td>
<td>35.4 (8.9)</td>
<td>0.3 (−0.7 to 1.2)</td>
<td>−7.8 to 8.4</td>
<td>8.1 (5.4 to 10.8)</td>
<td>0.90 (0.84 to 0.94)</td>
<td>2.9</td>
<td>0.95</td>
</tr>
<tr>
<td>WOSI (0 to 2100)</td>
<td>1055.7 (407.8)</td>
<td>1050.3 (444.6)</td>
<td>5.4 (−35.6 to 36.4)</td>
<td>−333.9 to 344.8</td>
<td>339.3 (227.0 to 451.8)</td>
<td>0.92 (0.87 to 0.95)</td>
<td>122.4</td>
<td>0.96</td>
</tr>
<tr>
<td>EQ-SD (&lt;−0.53 to 1)</td>
<td>0.70 (0.24)</td>
<td>0.71 (0.24)</td>
<td>−0.01 (−0.06 to 0.04)</td>
<td>−0.4 to 0.4</td>
<td>0.4 (0.3 to 0.5)</td>
<td>0.66 (0.50 to 0.77)</td>
<td>0.1</td>
<td>0.79</td>
</tr>
<tr>
<td>EQ-VAS (0 to 100)</td>
<td>71.9 (18.1)</td>
<td>72.3 (18.7)</td>
<td>−0.42 (−2.4 to 1.6)</td>
<td>−17.1 to 16.2</td>
<td>16.6 (11.2 to 22.2)</td>
<td>0.89 (0.83 to 0.93)</td>
<td>6.0</td>
<td>0.94</td>
</tr>
</tbody>
</table>

SD standard deviation, ICC (2.1) interclass correlation version 2.1 for measuring correlation between test and retest. Agreement estimated by the difference between test and retest, the limits of agreement (LoA), the standard error of measurement (SEM), and minimal detectable change (MDC) with 95% confidence interval. Chronbach’s alpha (internal consistency) are given for the 2.test. 95% CI (confidence interval) for paired t-test under null hypothesis = no difference between test and retest score.

† P<.0001 for all ICC (interclass correlation coefficient version 2.1).
Table 3 Agreement statistics, internal consistency and content validity for the domains of the WOSI 1. and 2.test

<table>
<thead>
<tr>
<th>Outcome (scores)</th>
<th>Median (min., max.)</th>
<th>Limits of agreement (LoA)</th>
<th>ICC (2.1)† (95% CI)</th>
<th>Floor effects %</th>
<th>Ceiling effects %</th>
<th>Minimal detectable change (MDC) (95% CI)</th>
<th>Standard error of measurement (SEMagreement)</th>
<th>Chronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical symptoms</td>
<td>452 (20, 897)</td>
<td>(~171.4 to 54.3)</td>
<td>0.92 (0.88 to 0.95)</td>
<td>0</td>
<td>0</td>
<td>162.9 (108.9 to 216.9)</td>
<td>58.8</td>
<td>0.96</td>
</tr>
<tr>
<td>Sports, recreation and work</td>
<td>243 (21, 398)</td>
<td>(~96.6 to 118.5)</td>
<td>0.82 (0.72 to 0.88)</td>
<td>0</td>
<td>0</td>
<td>107.6 (72.0 to 143.2)</td>
<td>38.8</td>
<td>0.90</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>190 (4, 399)</td>
<td>(~103.5 to 106.5)</td>
<td>0.87 (0.81 to 0.92)</td>
<td>0</td>
<td>0</td>
<td>105.0 (70.2 to 139.8)</td>
<td>37.8</td>
<td>0.93</td>
</tr>
<tr>
<td>Emotions</td>
<td>206 (2, 299)</td>
<td>(~91.2 to 116.2)</td>
<td>0.77 (0.65 to 0.85)</td>
<td>0</td>
<td>0</td>
<td>103.7 (69.3 to 138.1)</td>
<td>37.4</td>
<td>0.87</td>
</tr>
</tbody>
</table>

ICC (2.1), interclass correlation version 2.1 for measuring correlation between test and retest. Agreement estimated by the difference between test and retest, minimal detectable change (MDC) with 95% confidence interval, standard error of measurement (SEMagreement), and limits of agreement (LoA). Chronbach’s alpha (internal consistency) are given for the 2.test. Content validity is measured by floor and ceiling effects.

† 95% CI (confidence interval) for paired t-test under null hypothesis = no difference between test and retest score.

Table 4 Agreement statistics and content validity for the domains of the EQ-SD 1. and 2.test

<table>
<thead>
<tr>
<th>Outcome (scores)</th>
<th>Median (min. max.)</th>
<th>Limits of agreement (LoA)</th>
<th>Floor effects %</th>
<th>Ceiling effects %</th>
<th>Minimal detectable change (MDC) (95% CI)</th>
<th>ICC† (95% CI)</th>
<th>Standard error of measurement (SEMagreement)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>0.00 (0, 1)</td>
<td>(~0.30 to 0.35)</td>
<td>97.2</td>
<td>0.0</td>
<td>0.3 (0.2 to 0.4)</td>
<td>0.01 (~0.22 to 0.24)</td>
<td>0.2</td>
</tr>
<tr>
<td>Personal care</td>
<td>0.00 (0, 1)</td>
<td>(~0.63 to 0.61)</td>
<td>84.5</td>
<td>0.0</td>
<td>0.6 (0.4 to 0.8)</td>
<td>0.65 (0.49 to 0.77)</td>
<td>0.2</td>
</tr>
<tr>
<td>Daily activities</td>
<td>1.00 (0, 2)</td>
<td>(~0.85 to 1.04)</td>
<td>29.6</td>
<td>5.6</td>
<td>0.9 (0.6 to 1.2)</td>
<td>0.63 (0.47 to 0.75)</td>
<td>0.3</td>
</tr>
<tr>
<td>Pain/discomfort</td>
<td>1.00 (0, 2)</td>
<td>(~0.73 to 0.81)</td>
<td>22.5</td>
<td>9.9</td>
<td>0.8 (0.5 to 1.1)</td>
<td>0.73 (0.60 to 0.82)</td>
<td>0.3</td>
</tr>
<tr>
<td>Anxiety/depression</td>
<td>0.00 (0, 1)</td>
<td>(~0.66 to 0.57)</td>
<td>79.1</td>
<td>0.0</td>
<td>0.6 (0.4 to 0.8)</td>
<td>0.75 (0.63 to 0.84)</td>
<td>0.2</td>
</tr>
</tbody>
</table>

ICC (2.1), interclass correlation version 2.1 for measuring correlation between test and retest. Agreement estimated by the difference between test and retest, minimal detectable change (MDC) with 95% confidence interval, limits of agreement (LoA), and standard error of measurement (SEMagreement). Content validity is measured by floor and ceiling effects.

† p < 0.0001 for all ICC’s (interclass correlation coefficient version 2.1) except for the domain walking (p = 0.93).
substantially affect the size of the SEM. This inconsistency represents a problem for comparison between studies. SEM estimated as the square root of the mean square error term from the ANOVA avoids this problem, although the results will differ depending on the application of a one-way model or a two-way model as well as specification of fixed effects or random (individual) effects. The limits of agreement is not affected by the various methods used for calculating the ICC and SEM, and represents a uniform estimate of the measurement error that is easier to compare between studies. As shown in Tables 2 and 3, the limits of agreement were considerable for all questionnaires. For EQ-5D limits between -0.4 and 0.4 on a scale ranging from -0.53 to 1 means that this index is imprecise for estimating true change in an individual patient.

Content validity
In agreement with previous studies [2,6,12,37-43], all the questionnaires reflected the constructs to be measured. One of the aims of the present study was to evaluate the questionnaires for use in studies with patients with SLAP lesions, as the original versions of the OISS and WOSI were developed for use in patients with instability. Question 1 in OISS — During the last six months, how many times has your shoulder slipped out of joint (or dislocated)? — is not expected to be relevant for patients with superior labral tears (SLAP II lesions). However, 4 of 38 patients answered that their shoulder had slipped out of the joint, suggesting that they had the experience that this had occurred, or that they did not understand the question. Unfortunately,
we did not interview the patients about how they interpreted this question. The good content validity of the total scores of OISS and WOSI was supported by the absence of floor and ceiling effects for these questionnaires. Although single items of OISS had considerable floor and/or ceiling effects for both diagnostic groups, there were no floor or ceiling effects for single items of WOSI using the 15% definition. As noted by Ekeberg et al., agreement parameters can be overestimated when floor and ceiling effects appear, as an extreme value is more likely to be repeated in a retest [1]. The considerable floor effects of EQ-5D call into question the use of this generic self-report index in the population examined. The floor effects of EQ-5D suggest that health-related quality is not much affected by a SLAP-lesion or recurrent dislocation and that a specific questionnaire should be preferred. The use of EQ-5D cannot be recommended for use in cost-effectiveness studies in the present patient population. It may be better suited for shoulder patients who are expected to be more disabled, by example patients with comminute fractures of the humeral head [44].

Construct validity

In the present study, the construct validity was evaluated using both the correlation between instruments and the new criteria of the COSMIN group [32]. Previous studies have using correlation have reported good construct validity for OISS [6,20], WOSI [2,20,37-40], and EuroQol [21,22,24,41,45]. We found WOSI and OISS to be acceptably correlated for both diagnostic groups, which suggests that the self-report questionnaires can be used interchangeably. The EQ-5D, EQ-VAS, and Rowe score correlated < 0.60 with the specific questionnaires in both groups, which suggests that different constructs are measured. Applying the COSMIN checklist, OISS was acceptable for the two aspects of construct validity, but none of the questionnaires had acceptable convergent validity, but the use of hypotheses for the evaluation of construct validity is preferable, according to the COSMIN group [32] and to Guyatt [35]. The use of specific hypotheses also reduces the risk of bias, as stated by Terwee et al. [17], by avoiding the possibility of the retrospective construction of alternative explanations for the observed correlations. Nevertheless, the number of hypotheses applied can influence conclusions about validity.

Advantages and limitations of the study

The main advantages of the present study, in comparison with previous studies, are the evaluation of the scores according to recommendations in the COSMIN checklist. Although patients with SLAP lesions and patients with instability are comparable on most items, differences appeared [9]. One limitation of the current study is that the sample size of each diagnostic group is small; however no major differences appeared between groups. For future studies, including responsiveness, larger studies for each diagnostic group are recommended.

Conclusion

The measurement error and aspects of construct validity should be considered when OISS and WOSI are used in patients with recurrent shoulder dislocation and patients with SLAP-lesions. EQ-5D is not to be recommended as a single outcome instrument. The different methods for estimating SEM is a challenge when comparing measurement errors across studies.

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

ØS participated in the design of the study, drafted the manuscript, and evaluated patients for inclusion and follow-up exams. SL participated in the design of the study, presided the translation- and cross cultural adaptation process and helped to draft the manuscript. OR participated in the design of the study, contributed in monitoring the trial and drafting the manuscript. PM participated in the design of the study, planned and preformed the statistical analysis. JB participated in the design of the study, monitored the trial, contributed to the translation- and cross cultural adaptation process, and helped with drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Alastair Collins, Shari Gerber Nilsen, Susan Schanche for their contribution on the translation and cross cultural adaptation process of OISS and WOSI, and surgical director Lars Vasi, and colleagues and staff at Lovisenberg Diocional Hospital who allowed us to conduct this trial.

Author details

1Department of Orthopedic Surgery, Lovisenberg Diocional Hospital, Lovisenberggaten 17, 0440, Oslo, Norway. 2Department of Orthopedic Surgery, Sarungdset Hospital- Arendal, Sykehusveien 1, 4838, Arendal, Norway. 3Department of Orthopedic Surgery, Oslo University Hospital-Rikshospitalet, Sognsvannsvien, 0027, Oslo, Norway. 4Department of Pediatrics, Oslo University Hospital-Ullern, 0407, Oslo, Norway.

Received: 30 April 2013 Accepted: 8 July 2013
Published: 15 July 2013

References


Submit your next manuscript to BioMed Central
and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit
Efficacy of labral repair, biceps tenodesis, and diagnostic arthroscopy for SLAP Lesions of the shoulder: a randomised controlled trial

Øystein Skare1*, Cecilie Piene Schrøder1, Olav Reikerås2, Petter Mowinckel3, Jens Ivar Brox2

Abstract

Background: Surgery for type II SLAP (superior labral anterior posterior) lesions of the shoulder is a promising but unproven treatment. The procedures include labral repair or biceps tenodesis. Retrospective cohort studies have suggested that the benefits of tenodesis include pain relief and improved function, and higher patient satisfaction, which was reported in a prospective non-randomised study. There have been no completed randomised controlled trials of surgery for type II SLAP lesions. The aims of this participant and observer blinded randomised placebo-controlled trial are to compare the short-term (6 months) and long-term (2 years) efficacy of labral repair, biceps tenodesis, and placebo (diagnostic arthroscopy) for alleviating pain and improving function for type II SLAP lesions.

Methods/Design: A double-blind randomised controlled trial are performed using 120 patients, aged 18 to 60 years, with a history for type II SLAP lesions and clinical signs suggesting type II SLAP lesion, which were documented by MR arthrography and arthroscopy. Exclusion criteria include patients who have previously undergone operations for SLAP lesions or recurrent shoulder dislocations, and ruptures of the rotator cuff or biceps tendon. Outcomes will be assessed at baseline, three, six, 12, and 24 months. Primary outcome measures will be the clinical Rowe Score (1988-version) and the Western Ontario Instability Index (WOSI) at six and 24 months. Secondary outcome measures will include the Shoulder Instability Questionnaire (SIQ), the generic EuroQol (EQ-5 D and EQ-VAS), return to work and previous sports activity, complications, and the number of reoperations.

Discussion: The results of this trial will be of international importance and the results will be translatable into clinical practice.

Trial Registration: [ClinicalTrials.gov NCT00586742]

Background

The glenoid labrum contributes to stability by increasing joint concavity and dept of the glenohumeral joint socket. The superior glenoid labrum of the shoulder joint is a common site of injury and degeneration1,2,3]. Because it is related to the intraarticular insertion of the long head of the biceps tendon, injuries are common in throwing athletes. These lesions are often associated with other shoulder injuries such as rotator cuff tears, glenohumeral instability or impingement, but they also may be due to an isolated injury. Snyder et al. used the term SLAP (superior labrum anterior posterior) to describe these lesions, and they classified the lesions into four categories1]. Type II SLAP lesions, which occur most frequently, are characterised by the combined detachment of the superior labrum and biceps tendon from the peripheral edge of the glenoid. Surgical treatment includes reattachment of the labrum with the use of staples, metal screws, bioabsorbable tacks, and bioabsorbable anchors. Alternatively, tenodesis of the biceps tendon is performed, by inserting the tendon in the bicipital groove of the humeral head, either with suture anchors or interference screws.

Systematic reviews have analysed the value of diagnostic tests for SLAP-lesions3-7]. Recently, a systematic review summarised the current evidence about the outcome of...
type II SLAP repair]. Twelve studies, including 10 to 50 patients each, with at least 2-years of follow-up, were included; two studies compared two different surgical methods, two studies were prospective, while ten were retrospective cohort studies. There were no randomised trials. The percentage of patients classified as good to excellent varied from 40 to 94%. A return to their previous level of sports activity varied from 20 to 94%. Despite these unpredictable results and a lack of evidence from properly designed studies, shoulder surgeons worldwide perform type II SLAP repairs.

The aforementioned systematic review recommended that future studies should be prospective in nature and they should at least use a longitudinal prospective cohort design. Because uncontrolled studies have the potential to provide a distorted view of treatment results, and non-randomised trials are liable to produce biased results, we designed a prospective, randomised, double-blind, sham-controlled trial.

Aims
There are two aims of this randomised placebo-controlled trial:

1) Compare the short-term (6 months) efficacy of labral repair, biceps tenodesis, and placebo (diagnostic arthroscopy), for alleviating pain and improving function for type II SLAP lesions.

2) Compare the long-term (2 years) efficacy including the number of reoperations.

Methods/Design
Trial design
This is a participant and observer blinded randomised placebo-controlled trial with a 2-year follow-up (Figure 1).

Ethics
Ethics approval for this study has been received from the Ethics Committee Health Region Southeast, Oslo, Norway.

Participants
Participants will be recruited from general practitioners, physiotherapists, manual therapists, and from departments of orthopaedic surgery or physical medicine and rehabilitation. To increase the awareness of the trial, health care providers will be invited to attend lectures on shoulder complaints with a focus on the current study.

All potential participants will be screened to determine their eligibility according to the following inclusion and exclusion criteria. For inclusion, participants should be aged 18 to 60 years and have a history of type II SLAP lesions or clinical signs suggesting the presence of a type II SLAP lesion, and an MR arthrography that documents the type II SLAP lesion-. Finally, the diagnosis should be verified at arthroscopy. One experienced shoulder surgeon and one experienced manual therapist will perform clinical examinations of the patients. Participants should have at least one positive sign of a SLAP lesion (positive O’Brien test], positive Crank test], or painful apprehension test9]).

A thorough clinical examination will be performed to exclude possible candidates with differential diagnoses. The clinical examination will include tests for impingement10-], pain or weakness on isometric tests of abduction and external rotation], tests for apprehension and relocation], scapular dyskinesis], and arthritis of the acromioclavicular joint15]. Possible candidates will have an MR arthrography evaluated by a radiologist experienced in shoulder imaging. In addition, conventional x-rays including outlet view will be conducted to exclude patients with major acromioclavicular or acromial spurs.

Exclusion criteria include previous surgery for SLAP lesions, SLAP lesions with concomitant labral cysts16], previous surgery for recurrent shoulder dislocation or SLAP lesions, clinical and radiological signs of arthritis of the acromioclavicular15], or the glenohumeral joints, ruptures of the rotator cuff or biceps tendon11], synovial chondromatosis, fibromyalgia, major somatic or psychiatric disease, and patients that are not able to understand Norwegian or unwilling to accept one of the treatment alternatives.

Randomisation
Participants who fulfill the inclusion criteria, and consent to take part in the trial after they have received the oral and written information, will be randomised to receive labral repair, biceps tenodesis, or placebo (diagnostic arthroscopy) treatment. An independent statistician will use the method of permuted blocks for random allocation after the final inclusion criteria are met. Treatment allocation will be organised by an independent secretary who distributes sealed opaque numbered envelopes to the nurse manager in the operation theatre. A nurse will open the envelope only when a peroperative diagnostic evaluation has documented a type II SLAP lesion.

Interventions
The patient will be positioned in the lateral decubitus position with lateral traction and under general anaesthesia. A standard posterior portal will be created and a diagnostic evaluation will be performed. Prior to entering the glenohumeral joint the subacromial space will be inspected and evaluated. The subacromial and the glenohumeral evaluations will be documented in a video created for each patient. An anterior working portal will
be established in the rotator interval with a spinal needle for accurate placement. This portal will be used to probe the superior labrum for documentation of a type-II SLAP lesion. Arthroscopic diagnostic evaluations and treatments will be performed by a single experienced shoulder surgeon.

Following confirmation of a type II SLAP lesion, the patient will be included in the randomisation procedure. All patients will receive 20 to 40 ml of a 0.5% local anaesthetic (Marcaine) at the end of the procedure, partly to serve as a suprascapular nerve block and partly to serve as an intraarticular injection. A collar and cuff sling will be placed before the patient leaves the operating room. **Placebo (diagnostic arthroscopy)** Patients randomised to diagnostic arthroscopy and postoperative rehabilitation will comprise the placebo group. **Labral repair** Debridement of the superior glenoid rim will be performed with a motorized shaver from the anterior portal. The bioabsorbable suture anchor will be placed percutaneously, guided by a spinal needle through the myotendinous junction of the supraspinatus. From the percutaneous portal two suture anchors will be placed.
in the glenoid posterior to the insertion of the biceps tendon. Sutures will then be made with the use of a shuttling device from the anterior portal. Fixation will be secured with a sliding knot and three half-hitches in alternating directions. Eventually, an anterior anchor will be placed through the anterior portal. No other procedures will be performed.

**Biceps tenodesis**

Although other arthroscopic methods are described, we routinely use a mini-open technique for biceps tenodesis (14). For exact positioning of the biceps tendon, a spinal needle will be placed under arthroscopic vision, as far laterally and central as possible in the biceps tendon. A tenotomy will be performed at the biceps labrum junction. The rest of the procedure will be performed mini-open with a 2 cm skin incision with the spinal needle in the centre. In order to identify and open the biceps pulley the deltoid will be split along the muscle fibers. The biceps tendon will be identified and lifted outside of the bicipital groove. The groove will be debrided, and a metal double suture anchor with needles will be placed in the groove. One of the limbs of each suture will be placed as a simple stitch to secure sliding of the knot, and the second limb will be passed two times to secure the fixation. Approximately 2 cm of the tendon will be excised and the pulley and skin will be closed. No other procedures will be performed.

**Post-operative rehabilitation**

Patients in all three groups will have standardised, but individually adjusted rehabilitation. Elbow, wrist, and finger mobilisation and gentle pendulum exercises will be conducted, starting on the first postoperative day. A sling will be used for three weeks. Local physiotherapists or manual therapists, who are given a written detailed description of the methods and progression, will provide treatment to patients when they are discharged from the hospital. Passive techniques like massage and stretching along with core stability exercises and general physical training will be used during the first three weeks. Exercises to normalise the gleno-humeral rhythm and improve coordination and mobility will be given using sling exercise therapy17]. Exercises to improve functional stability and muscle strength of the rotator cuff and scapular stabilising muscles will be progressively emphasised after six weeks. Sports- or job-specific rehabilitation will be given on an individual basis, usually starting three months postoperatively. Rehabilitation will continue for three to six months and will include 12-16 sessions with a therapist and about 20 sessions of self-administered exercises.

**Outcome assessment**

Baseline data will include gender, age, smoking, previous treatment, duration of symptoms, MR arthrography and conventional x-rays including outlet view, and primary and secondary outcome measures.

The same blinded observer will assess all participants after the procedure at three, six, 12 and 24 months. Pain, health related quality of life, complications, and a return to sports and work will be assessed at each time point. Blinding will be evaluated by asking the patients about which treatment they perceive to have received.

Pain during activity and pain at rest (over the last week) will be measured on a 0–100 visual analogue scale (VAS), comprising a horizontal line labelled no pain at one end and worst imaginable pain at the other end.

A range of standardised, generic and specific self-report health-related quality of life measures and the clinical Rowe Score will be used. To our knowledge outcome measures have not been particularly evaluated for patients with SLAP lesions. The primary outcome measures in the present trial will be the 1988 version of the Rowe Score18,] and the Western Ontario Instability Index (WOSI)19. The latter has been professionally translated to Norwegian.

The Rowe Score was first described in 1978 for use in patients after they were administered the Bankart procedure for anterior shoulder dislocation20]. Four different versions exist. We will use the 1988 version. The observer will question the patient about function and pain, and assess their stability, muscle strength, and range of motion. The Rowe Score can be weighted using either pain or stability as the main problem. Because pain is the main complaint in patients with type II SLAP lesions, we will weight pain as 25 points. Pain has five levels ranging from severe (0 points) to none (25). Stability has five levels ranging from recurrent dislocation (0) to normal shoulder stability, which includes a negative apprehension test (15). Function has five response alternatives from total disability (0) to normal function with no limitation in daily living, sports, or work (25). Range of motion is evaluated for abduction/forward flexion, internal rotation and external rotation, and it is categorised from a full range of motion (25) to less than 30° of motion (0). Muscle strength will be measured by a spring gauge, and results will be compared to the opposite shoulder and categorised from normal (10) to poor (0). The best achievable score is 100. Results are commonly classified into four categories: poor (39 points or less), fair (40 to 69 points), good (70 to 89 points), and excellent (90 to 100 points).

The WOSI is a disease-specific health related quality of life instrument developed and validated for use in patients with shoulder instability. It comprises 21 items representing four domains. The first domain covers physical symptoms and contains 10 items. The remaining domains are sports, recreation, and work (four items), lifestyle (four items), and emotions (three items). Each
question is scored from 0 (best possible) to 100 on a visual analogue scale. The worst score possible is 2100. This signifies that the patient has an extreme decrease in shoulder-related health-related quality of life.

The Shoulder Instability Questionnaire (SIQ) is a disease-specific health related quality of life instrument validated for use in patients with shoulder instability. It includes 12 questions (1-5 points each) with possible scores from 12 (best function) to 60 (worst function).

The EuroQuol (EQ-5 D and EQ-VAS) is a standard generic health-related quality of life instrument. The EQ-5 D measures five domains (Mobility, Self-Care,Usual Activities, Pain/Discomfort, and Anxiety/Depression); each has three levels, ranging in severity from no problem, to some problem, or an extreme problem. Responses are transformed to an index and then classified into 243 (3^8) health states, with the best imaginable state (1.0) representing the highest level of functionality.

Sickness absence data will be collected from the National Social Security Institution.

Sample size
The main end-points are six and 24 months. From clinical experience we estimated that the smallest clinically important detectable difference is 10 points on the 100 points Rowe Score. Assuming that the largest difference between treatments will be 10 units, we simulated multiple scenarios and estimated the standard deviation between means to be 14.6 units. To detect this difference between treatment groups (SD = 15, α = 0.05, β = 0.80, One-Way ANOVA) our study will require 36 patients in each group. Assuming some patients dropout, we plan to include 40 patients in each group.

Planned statistical analysis
Treatment groups will be examined for comparability at baseline with respect to demographic and prognostic factors. All eligible patients, regardless of their compliance with protocol (analysis by intention-to-treat) will be included in the main analyses. To assess the effect of the interventions on the endpoints (six and 24 months), analysis of covariance (ANCOVA) will be performed using the baseline values as one of the covariates. Standard regression assumptions will be assessed using diagnostic plots, Jackknife residuals, Cook’s distances, and Variance inflation Factor (VIF). We will adjust for an eventual imbalance at the baseline. Corresponding post-hoc tests (Tukey’s test) will be performed. To evaluate the time-course at three, six, 12, and 24 months, repeated measures will be analyzed using linear mixed models. If the number of missing values exceeds 10% in one of the groups, multiple imputations will be used to estimate the missing values. To assess the robustness of our findings the analysis will be performed with and without the imputed values.

Discussion and conclusion
Surgery for type II SLAP lesions are performed worldwide, but published reports suggest that outcome is difficult to predict. Interventions that effectively reduce pain, improve function, and allow patients to return to sports and work are lacking. Promising results are published for both biceps tenodesis and labral repair, but the lack of a randomised design, standardised inclusion and exclusion criteria, and small study sizes, may bias these conclusions.

Few clinical trials in orthopaedic surgery include sham or placebo treatments. Two trials compared vertebroplasty with placebo in patients with osteoporotic vertebral compression fractures, and one trial compared arthroscopic lavage, debridement, and placebo in patients with osteoarthritis of the knee. Neither of these trials found that the surgical procedure was effective compared with the placebo. These trials emphasise the importance of including a placebo intervention in a randomised trial in order to improve present knowledge about mechanisms for pain reduction after surgical procedures.

Acknowledgements
We would like to thank the surgical director Lars Vasli and colleagues and staff at Lovisenberg Deakonal Hospital who allowed us to conduct this trial.

Author details
1. Lovisenberg Deakonal Hospital, Lovisenberggaten 17, 0440 Oslo, Norway.
2. Orthopaedic Department, Oslo University Hospital-Rikshospitalet, Sognsvannsveien, 0227 Oslo, Norway, and University of Oslo. 3. Pediatric Department, Oslo University Hospital-Ullevål, 0407 Oslo, Norway.

Authors’ contributions
ØS participated in the design of the study, drafted the manuscript, and evaluated patients for inclusion and follow-up exams. CPS participated in the design of the study, operated on all patients included in the study, and planned and performed the statistical analyses. JIB participated in the design of the study, drafted the manuscript, and monitored the trial. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 28 May 2010 Accepted: 7 October 2010
Published: 7 October 2010

References