MASTER OF MEDICINE

ORTHOPAEDIC SURGERY

RELIABILITY OF SHOULDER SYMPTOM RECALL AFTER ONE YEAR IN A RETROSPECTIVE APPLICATION OF THE OXFORD SHOULDER SCORE

BY

MICHAEL HELD

DR. MED.

ORTHOPAEDIC REGISTRAR

GROOTE SCHUUR HOSPITAL

UNIVERSITY OF CAPE TOWN

SUPERVISORS: PROF. J. WALTERS, DR. S. ROCHE

Reliability of shoulder symptom recall after 1 year in a retrospective application of the Oxford Shoulder Score

ABSTRACT

Background

The accuracy of retrospective recall of shoulder symptoms has not been well documented. This prospective study assesses the ability of patients to recall their preoperative shoulder function one year after a surgical intervention, using the Oxford Shoulder Score (OSS).

Methods

35 patients completed an OSS before undergoing shoulder surgery. One year later, patients were asked to recall their symptoms prior to their surgery. The recalled OSS of the patients as a group was compared to their preoperative score. The recall bias of each test pair was assessed with a Bland – Altman plot.

Results

On recall after a mean of 12.6 months, the mean OSS from the index assessment increased from 36.25 to 38.25 points. The mean difference of 2 points for the patients as a group was not significant ($p = 0.14$). The statistical limits of agreement of the Bland – Altman plot were set at +/– 2 SD = 14.079 points. The plotted points showed fair correlation between each individual test pair.

Conclusion

The recall of symptoms of a large group of patients at 1 year after the index intervention appears to have a moderate correlation with the preoperative scoring. Although statistically acceptable, this limit of agreement is much larger than the 4.5-point difference, established to be clinically relevant in prior studies. The variation seen within the scores at the individual level suggests that these data cannot be used as a retrospective tool.

CONTENTS

P U B L I C A T I O N – R E A D Y F O R M A T

Part A

RESEARCH PROTOCOL

Reliability of shoulder symptom recall after one year in a retrospective application of the Oxford Shoulder Score

I. Background

The accurate assessment of the functional status and symptoms experienced by patients is a central component of clinical evaluation and fundamental to well-designed research. Quantitative scoring systems, as well as patient reported outcome measures such as the Oxford Shoulder Score (OSS), are tools to enhance this process. (Dawson et al., 1996)

Repeated measures, where individuals are scored before and after an operative procedure, have exceptional statistical power. However, when the cause of a shoulder condition is acute, or a preoperative assessment could not be obtained for other reasons, answering questions on pre-injury symptoms and functional status becomes a retrospective implementation.

II. Purpose

The aim of this study is to investigate the reliability of the OSS when applied retrospectively to assess shoulder symptom recall. The most valuable implication is to assess and quantify individual outcome and postoperative functional improvement.

III. Definition of terms

OSS: Oxford Shoulder Score

Arthroplasty: Replacement of the shoulder joint

Subacromial decompression: Partial removal of bone on the under surface of the acromial process of the scapula

Acromio-clavicular joint excision: Removal of the joint between the acromion and the clavicle

Rotator cuff procedures: Interventions involving the muscular cuff around the shoulder joint.

IV. Methods

This prospective study will include 35 adult patients attending shoulder surgery under the care of one dedicated shoulder surgeon. Patients will be asked to complete an OSS to document their current symptom and functional level. The interventions will include arthroplasty, arthroscopic subacromial decompression, acromio-clavicular joint excision, and rotator cuff procedures. These procedures as well as the postoperative rehabilitation will take place at Vincent Pallotti Hospital, Groote Schuur Hospital and Constantiaberg Medi-Clinic in Cape Town.

Children under the age of 18 years, adults with impaired decision-making capacity, people highly dependent on medical care, people with unequal or dependent relationships will be excluded from this study.

The OSS (Dawson et al., 1996) is a validated, 12-item scoring-tool specifically designed for patients having shoulder operations, not including stabilization. (Constant and Murley, 1987; Fries, 1982) It is purely patient-based, with each item graded from 1 to 5. Scores are added to give a single total. The minimum score is 12 (best), maximum 60 (worst). Four items relate to pain, the other 8 relate to ability to perform everyday tasks with the affected upper limb.

The scores will be carried out 24 hours prior to surgery. At 6-12 months after the procedure, the patients will be contacted telephonically (~15 minutes) by a dedicated orthopaedic surgeon or trained health-care worker and will be asked to complete the OSS again. They will be asked to remember what their symptoms were during the week before their surgery. The preoperative and recall score will be compared to each other. Data will be analysed using SPSS 13.0 (SPSS Inc.)

V. Description of risks and benefits

The patients will be treated routinely with no change to the surgeon's usual protocols as per internationally accepted standard of care. The risks and benefits are identical for patients not included in this study.

VI. Ethical considerations

Informed consent - Informed consent will be taken telephonically and an introduction of the study given, prior to the assessment of the OSS. The patients will

have the options to (1) participate, (2) to be approached at a later stage after sufficient time to consider a possible participation or (3) to reject participation. This consent will be taken by means of a standard consent script as below.

Data safety and reimbursement - All patient names and folder numbers will be removed from the data stream. This study adheres to the Declaration of Helsinki 2008. There will be no reimbursement.

VII. Researchers

Michael Held, Stephen Roche, Basil Vrettos, Maritz Laubscher, Johan Walters
Orthopaedic Department, Groote Schuur Hospital, University of Cape Town

Part B

LITERATURE REVIEW

I. Objectives

a. To discuss the significance of patient-based outcome scores in Orthopaedic Surgery.
b. To establish an understanding of scoring and assessing in Shoulder surgery.
c. To review the importance and clinical relevance of the Oxford Shoulder Score (OSS).
d. To identify similar studies assessing retrospective recall of the OSS.
e. To evaluate methods to limit bias when comparing two scores with each other.

II. Summary and interpretation of literature for the research

Outcome measures as tools to track and acknowledge errors

"Real knowledge is to know the extent of one's ignorance" (Confucius). This concept was addressed by Ernest Amory Codman and his Back-Bay-Golden-Goose-Ostrich (Figure 1) in 1915 as it buries its head in the sand and ignores what is happening around it. As an attempt to promote his concept of the "End Result System" he satirized his colleagues to whom it was a foreign concept to practice medicine by tracking and acknowledging errors.

Figure 1. Codman's cartoon of an ostrich burying its head in the sand, symbolising medicine practiced without the feedback of outcome measures and tracking of errors made.

Unfortunately, his peers did not share the same enthusiasm and his famous cartoon stands as evidence of his lack of diplomacy, which alienated his peers. As in many

other promoters of ideas, which were ahead of the time, Codman's career declined thereafter and he died in relative anonymity. Yet, he sparked the pursuit of 'best practice', which led to today's evidenced-based strategies in managing our patients.

Types and application of outcome measures in Shoulder surgery

Even though the importance of tracking results and thereby learning from mistakes has been accepted by now, scoring and appropriate use of certain scores is poorly understood. The two broad groups of outcome measures are general health scores (i.e. Medical Outcomes Study 36-Item Short Form, SF-36) and joint-specific or disease-specific scores (i.e. Oxford Shoulder Score). To compare specific scores of one population to another, it is important to include a general health measure in the study protocol.

To date, more than 30 different outcome scores have been developed to assess shoulder symptoms, yet many are being applied inappropriately. This leads to flawed conclusions of research projects, difficulties with comparing trials with each other and, most importantly, to limitations in assessing patient recovery and clinical outcome. Harvie et al. have assessed the manner of application of frequently encountered outcome scores (Figure 2) and have stated that most scores are applied inappropriately. They see modification from the scores' original use and improper testing for validity, repeatability and sensitivity to change as the main culprits. (Harvie et al., 2005) In other words, looking at Codman's ostrich 100 years later, our head is still in the sand, but at least we are aware of it.

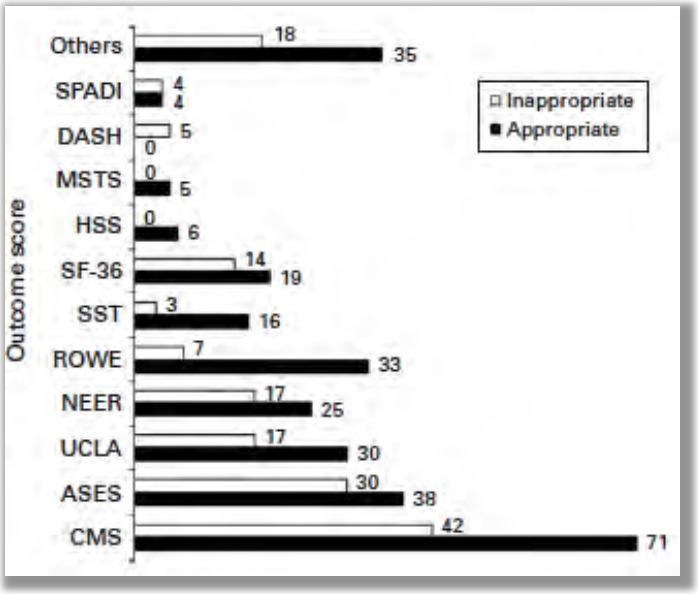

Figure 2. Manner of application of frequently encountered outcome scores as described by Harvie et al. (CMS, Constant-Murley shoulder score; ASES, American Shoulder and Elbow Surgeons standardised shoulder assessment form; UCLA, University of California Los Angeles shoulder rating scale; Neer, Neer shoulder rating; Rowe, Rowe instability score; SST, simple shoulder test; SF-36, 36-item short-form health survey; HSS, hospital of Special Surgery shoulder assessment; MSTS, Musculo-skeletal tumour score; DASH, Disabilities of the arm shoulder and hand questionnaire; SPADI, shoulder pain and disability index).

Various expert bodies, such as The European Society for Surgery of the Shoulder and the American Academy of Orthopaedic Surgeons (AAOS), have attempted to give guidance on the appropriate application of outcome scores, yet even those recommendations are not uniformly accepted. Below is a list of the currently available shoulder scores and their preferred use (Table 1):

<u>General, Arthroplasty, Osteoarthritis</u>
American Shoulder and Elbow Surgeons Shoulder Outcome Score (ASES)
Constant Score
Croft measurement of shoulder-related disability
(I.e., The Disability Questionnaire, United Kingdom Shoulder Disability Questionnaire)
Disabilities of the Arm, Shoulder, and Hand (DASH)
Flexilevel Scale of Shoulder Function (FLEX-SF)
Hospital for Special Surgery Score (HSS)
L'Insalata Shoulder Rating Questionnaire (SRQ)
Neer Shoulder Score
Oxford Shoulder Score
Penn Shoulder Score
Shoulder Activity Level
Shoulder Disability Questionnaire–Dutch (SDQ-NL)
Shoulder Pain and Disability Index (SPADI)
Shoulder Pain Score
Shoulder Range of Motion Questionnaire
Simple Shoulder Test (SST)
Single Assessment Numeric Evaluation (SANE)
Subjective Shoulder Rating System (SSRS)
University of California Los Angeles (UCLA) Shoulder Score
Western Ontario Osteoarthritis of the Shoulder Index (WOOS)
<u>Rotator Cuff Disease</u>
Rotator Cuff Quality of Life (RCQOL)
Western Ontario Rotator Cuff Index (WORC)
Wolfgang criteria for rating results of rotator cuff surgical repair Continued →

Shoulder Instability
Rowe Rating Sheet for Bankart Repair
Western Ontario Shoulder Instability Index (WOSI)
Oxford Instability Score (OSS-I)

Table 1. Currently available scores and their preferred use

The recommendation of the AAOS to use these scores in research is as follows: (Wright and Baumgarten, 2010)

1. Include a general health outcome measure to be able to compare studies to other populations (SF-36).
2. Include activity level measures to compare outcomes among patients as low activity level can falsely elevate many outcome scores. (Brophy et al., 2005)
3. Choose a general shoulder score or a specific shoulder score as indicated in Table 2.

Indication	Score	Reason for use	Optional Scores
Compare different diagnoses/cross-sectional/general shoulder measure for research purposes	ASES	Popularity	OSS
Evaluate various diagnoses quickly	ASES, UCLA		
Workman's compensation claims	DASH		
Rotator cuff	WORC, RCQOL	Established MCID	
Shoulder instability	WOSI, OSS-I		ROWE, if comparing to older studies
Arthritis/Arthroplasty	WOOS		UCLA/Neer, if comparing to older studies

Table 2. The recommendation of the AAOS to use scores according to the diagnosis: American Shoulder and Elbow Surgeons shoulder outcome score (ASES), Disabilities of the Arm, Shoulder, and Hand (DASH), Neer shoulder score (Neer), Oxford shoulder score (OSS), University of California Los Angeles (UCLA) Shoulder Score, Western Ontario Osteoarthritis of the Shoulder index (WOOS), Rotator Cuff Quality of Life (RCQOL), Western Ontario Rotator Cuff Index (WORC), Rowe Rating, Sheet for Bankart Repair (ROWE), Western Ontario Shoulder Instability Index (WOSI), Oxford Instability Score (OSS-I).

Patient-reported outcome measures (PROMs)

These recommendations are aimed at research and will help with designing sound study protocols, but each clinical setting has to be assessed individually to weigh the information gained with these recommended scores against the burden on patients and surgeons to complete the scores. With this vast number of outcome scores in Shoulder surgery, certain criteria can be used to evaluate whether they are suitable for the specific research question or clinical setting. Scale development, appropriateness, reliability, validity, responsiveness, precision, interpretability, acceptability and feasibility should be evaluated (Fitzpatrick et al., 1998).

Unfortunately, we are often faced with poor documentation from medical staff and dysfunctional filing systems. The contact time of patient and surgeon is reduced to a minimum and time and resources are limiting factors when outcome scores are collected. Although this might not leave room for excessive scoring of patients, it remains critical, even with these limitations encountered to assess and improve patient management. Patient-reported outcome measures (PROMs) seem to play a promising part to overcome these limitations, especially in the absence of a research question.

PROMs describe illness and benefits of health care interventions from the patient's perspective and can be divided into 3 concepts (Valderas and Alonso, 2008): (1) construct (the measurement object); (2) population (based on age, gender, condition, and culture); and (3) measurement model (dimensionality, metric, and adaptability). Most PROMs assess more than one of these constructs. One of the main advantages of PROMs is that patients can complete these outcome measures prior to the assessment by the surgeon. This saves time and allows to direct the clinical shoulder assessment towards specific points raised in the outcome measure by the patient.

This could be the reason why over the last 20 years the use of patient-reported outcome measures (PROMs) has increased significantly (Figure 4).(Harvie et al., 2005)

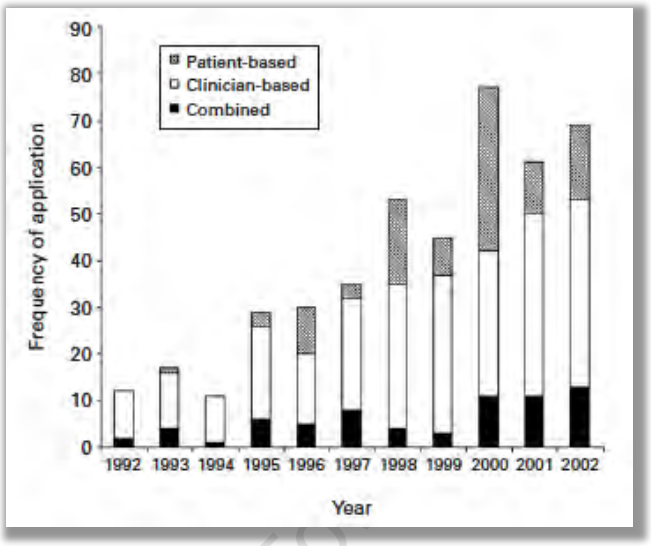

Figure 4. The frequency of patient-based outcome scores has increased over the last 20 years, whereas clinician-based scoring has remained constant. (Harvie et al., 2005)

Yet, even with the use of PROMs, certain considerations need to be taken into account, as poor data collection cannot be restored with statistical methods at a later stage (Dawson et al., 2010):

1. Flawed data cannot be improved through analysis. Collaboration with statisticians from the beginning of data collection is crucial.
2. Specifying the reason for data collection with reference to an event/intervention and defining the follow-up period should be done before data collection is started.
3. It is important to ensure that cross-sectional and longitudinal records of each patient are accurately linked, by using a unique patient identifier.
4. For bilateral structures (joints, limbs), deciding in advance on the unit of analysis (e.g. right versus left) will avoid errors especially with multiple measurements.
5. Scores need to be recorded with the date of completion (not the date of data entry), with reference (linked or labelled) to the date of an intervention.
6. Avoiding duplication of entries can be done by setting up automatic prompts.
7. Using a pilot to assess and review a systematic method of data collection will point out pitfalls. Once a method has been devised it should be adhered to.
8. The data in spreadsheets or database should be clearly labelled to facilitate statistical analysis without complex data programming.

9. Conducting simple analyses and downloading data should be checked when no more than 20 cases have been entered.
10. Great effort should be made to complete follow-up data. The intensity of which follow-up information is sought, greatly influences the response rate.

The Oxford Shoulder Score (OSS)

As a tool of PROMs, the Oxford Shoulder Score has been rigorously validated and studied. It was developed as a joint specific instrument assessing the outcomes in shoulder surgery and avoiding potential reporting bias of surgeons assessing their own patients' improvement. (Dawson et al., 2009) It is completed by the patient and is a 12-item score; each item is graded from 1 to 5 adding to a single total score. The best score is 12 and the worst is 60. Pain is assessed through four items. Activities of daily living are assessed though the remaining 8 items. It is sensitive to change and has good test-retest reliability at 24 hours. The smallest amount of change on a measure, which patients perceive to be of clinical importance still needs to be determined for the OSS. For many PROMs this number is about half of the standard deviation of change. (Dawson et al., 2009)

Recall reliability of the Oxford Shoulder Score

Unfortunately proper systems for data capture and storage of outcome measures are often not in place. This problem is fuelled by a large number of trauma patients in which prospective scoring is impossible, but also due to poor documentation from the medical staff and problems of data retrieval from dysfunctional filing systems. Despite these circumstances, we would be able to compare preoperative and postoperative shoulder function in patients if they could remember their symptoms before their surgery, even at a later stage of their postoperative evaluation.

In a previous study, Wilson et al. (Wilson et al., 2009) evaluated the reliability of shoulder symptom recall and the conclusion was drawn that it seems valid to assess the impact of a shoulder operation in a large group of patients within a population (Figure 5). The mean time to recall after the surgery was 50 days (range 22-150). The mean age of the 63 patients included was 57 (range 20-86). Wilson also compared patients under 50 with patients over 60 years of age and found that age did not adversely influence the recall.

However, based on the BA plot (Figure 6), the authors found poor correlation when comparing the individual scores of each patient. Outliers exceeded the limits of agreement they set at a 4,5 point difference, which is considered clinically relevant by Cloke et al. (Cloke et al., 2008)

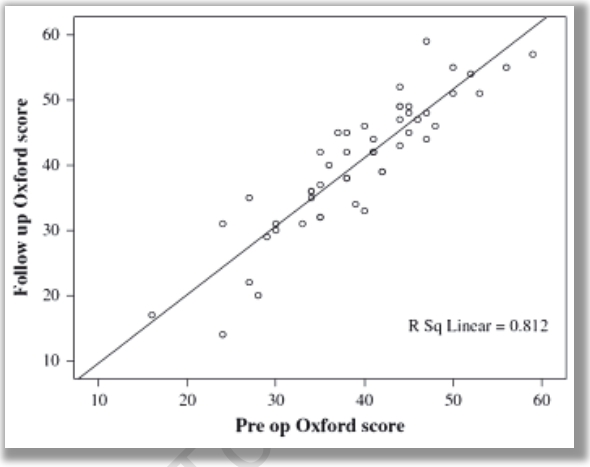

Figure 5. Scatter plot of the study of Wilson et al., showing the correlation
of the pre-operative Oxford Shoulder Score compared to the score on follow-up, when remembering previous symptoms.

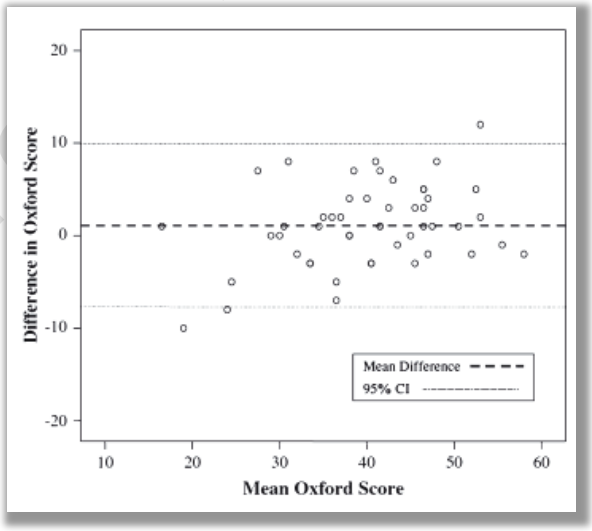

Figure 6. Bland–Altman plot of the study of Wilson et al., showing the
difference of each individual test pair plotted against the mean Oxford
Shoulder Score. Outliers exceeded the limits of agreement they set at a 4,5 point difference, which is considered clinically relevant by Cloke et al. (Cloke et al., 2008)

Specific statistical methods – Bland Altman plot

In this study, two scores are compared to each other. Initially, patients will be asked to complete the Oxford Shoulder Score preoperatively. One year later, they will be asked if they can remember what their symptoms were initially and the recalled score will be documented and compared to the initial score. The Bland-Altman plot was used to assess the repeatability of the Oxford Shoulder Score at two time points and to show that the variability is not related to the magnitude of the score.

A simple example will illustrate the importance of a method designed to assess bias when comparing two different groups of data.

Group A:	
Score I:	5
Score II:	25
Total Score of Group A:	30
Mean score of Group A:	15 (30 : 2)

Group B:	
Score I:	15
Score II:	15
Total Score of Group B:	30
Mean score of Group B:	15 (30 : 2)

As evident in the two groups above, the mean score of Group A is identical to the mean score of Group B. Yet the individual Test scores differ. Bland and Altman developed a plot to show bias when comparing two different tests with each other. (Bland and Altman, 1986) The difference of each test pair is plotted against their mean. The 0-line indicates a difference of 0 points; hence, two pairs produce equal results. Statistically, the limits of agreement are defined as the mean difference +- 2 SD of the differences, which indicates that 95% of the values are normally distributed.

Summary

Tracking and acknowledging errors through outcome measures is fundamental to evidenced-based strategies in the management of surgical patients. Challenges arise when the appropriate application of scores and evaluation processes have to be balanced with their feasibility within our clinical setting. Keeping this in mind, patient-based outcome scores are practical, save time and can provide valuable information on illnesses and benefits of health care interventions from the patient's perspective. The Oxford Shoulder Score is a patient-based outcome measure, which has been rigorously validated and studied. It is a 12-item scoring tool assessing pain and activities of daily living in patients undergoing shoulder surgery. Using this test, the reliability of remembering symptoms and shoulder function would enable us to compare the preoperative to postoperative shoulder function in patients, even at a

later stage during their postoperative evaluation. This would be essential to deal with inadequate preoperative documentation, encountered in our setting. A previous study by Wilson has reported an overall significant correlation between pre- and postoperative assessments of a cohort of patients conducted within two months. However, they recorded a wide variation in the assessments of individuals. (Wilson et al., 2009).

III. Identification of gaps or needs for further research

Our study had some gaps and weaknesses. Only one assessment tool, namely the Oxford Shoulder score, was used. Other scoring tools with a more general assessment (i.e. SF-36) would have allowed our study to be compared to other patient populations. In our study the interval between the two tests was one year, compared to 6 weeks in the study of Wilson et al., thereby eliminating one of the weaknesses Wilson found in their study. The longer time period in our study seems more appropriate as it allowed for complete healing and rehabilitation after the operation. Thus, patients would have to recall their symptoms, rather than being reminded by residual pain or functional deficits.
We concluded that our patients cannot recall the severity of their preoperative shoulder symptoms reliably and we might need to reassess how much relevance we put on comparing them to their postoperative scores. In future studies we should evaluate at what point in time after the operation a patient cannot recall his preoperative shoulder symptoms anymore, as this might be the point when, subjectively, recovery takes place. It is possible that these findings will be applicable to clinical areas beyond the field of orthopaedic surgery, which rely on patient based clinical outcome measures.

IV. Literature search strategy

Literature search was done online. Search engines, such as Google scholar (www.scholar.google.com) and PubMed (www.pubmed.gov) were used. The following key words were used: Oxford shoulder score, patient based outcome measures, reliability of recall of symptoms, shoulder scores. Articles before the year of 1960 were excluded.

V. Quality criteria

To ensure the quality of this study the following points were raised:

a. Context for the study

We describe a clear context for the study: Often, preoperative scoring is done insufficiently or data is lost in deficient filing systems, thus if patients could remember their preoperative shoulder function, this would allow for comparison to postoperative scores.

b. Aim of the study

The aim of the study is clearly defined. Time margins (1 year) as well as the research tool (Oxford Shoulder Score) are stated.

c. Data collection and analysis

The data analysis is described and certain methods are used, such as the Bland Altman plot (Bland and Altman, 1986), to avoid bias when comparing the data. Clinical importance of the data was reassessed and limits of agreement for clinical relevance was adjusted to avoid misinterpretation of data, which was derived from the Bland and Altman plot. The time interval between the two tests was one year and allowed for complete healing and rehabilitation after the operation. Thus, patients would have to recall their symptoms, rather than being reminded by residual pain or functional deficits.

d. Conclusions supported by the results

We show limited support for our initial question: Can patients remember their shoulder function? This is only true for a large group of patients, not for each individual.

e. Validity and reliability

Our research tool is the Oxford Shoulder Score, which has been extensively studied and validated. (Dawson et al., 2009)

f. Replicability

This study has good replicability since comparable interpretations and conclusions could be drawn from our study as from a similar study by Wilson. (Wilson et al., 2009)

VI. References

BLAND, J. M. & ALTMAN, D. G. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet,* 1**,** 307-10.

BROPHY, R. H., BEAUVAIS, R. L., JONES, E. C., CORDASCO, F. A. & MARX, R. G. 2005. Measurement of shoulder activity level. *Clin Orthop Relat Res,* 439**,** 101-8.

CLOKE, D. J., WATSON, H., PURDY, S., STEEN, I. N. & WILLIAMS, J. R. 2008. A pilot randomized, controlled trial of treatment for painful arc of the shoulder. *J Shoulder Elbow Surg,* 17**,** 17S-21S.

CONSTANT C.R., Murley, A.H.G. 1987. A clinical method of functional assessment of the shoulder. *Clin Orthop*, **214**, 160–164.

DAWSON, J., ROGERS, K., DOLL, H., FITZPATRICK, R., COOPER, C. & CARR, A. J. 2010. Using Patient-Reported Oucome Measures (PROMs) Routinely: Example in the Context of Elective Shoulder Surgery. *The Open epidemiology Journal,* 3**,** 10.

DAWSON, J., ROGERS, K., FITZPATRICK, R. & CARR, A. 2009. The Oxford shoulder score revisited. *Arch Orthop Trauma Surg,* 129**,** 119-23.

FITZPATRICK, R., SHORTALL, E., SCULPHER, M., MURRAY, D., MORRIS, R., LODGE, M., DAWSON, J., CARR, A., BRITTON, A. & BRIGGS, A. 1998. Primary total hip replacement surgery: a systematic review of outcomes and modelling of cost-effectiveness associated with different prostheses. *Health Technol Assess,* 2**,** 1-64.

FRIES J.F., Spitz P.W., Young D.Y. 1982. The dimensions of health outcomes: the Health Assessment Questionnaire, disability and pain scales, *J Rheumatol*, **9**, 789–793.

HARVIE, P., POLLARD, T. C., CHENNAGIRI, R. J. & CARR, A. J. 2005. The use of outcome scores in surgery of the shoulder. *J Bone Joint Surg Br,* 87**,** 151-4.

VALDERAS, J. M. & ALONSO, J. 2008. Patient reported outcome measures: a model-based classification system for research and clinical practice. *Qual Life Res,* 17**,** 1125-35.

WILSON, J., BAKER, P. & RANGAN, A. 2009. Is retrospective application of the Oxford Shoulder Score valid? *J Shoulder Elbow Surg,* 18**,** 577-80.

WRIGHT, R. W. & BAUMGARTEN, K. M. 2010. Shoulder outcomes measures. *J Am Acad Orthop Surg,* 18**,** 436-44.

References used in the manuscript not included in the literature review:

BAKER P, Nanda R, Goodchild L, Finn P, Rangan A. 2008. A comparison of the Constant and Oxford shoulder scores in patients with conservatively treated proximal humeral fractures. J Shoulder Elbow Surg. 17(1):37-4

CLOKE DJ, Watson H, Purdy S, Steen IN, Williams JR. 2008. A pilot randomized, controlled trial of treatment for painful arc of the shoulder. J Shoulder Elbow Surg, 17(1 Suppl):17S-21S.

DAWSON J, Carr A. 2001. Outcomes evaluation in orthopaedics. J Bone Joint Surg Br, 83(3):313-315.

DAWSON J, Fitzpatrick R, Carr A. 1996. Questionnaire on the perceptions of patients about shoulder surgery. J Bone Joint Surg Br, 78(4):593-600.

GERMANN G, Wind G, Harth A. 1999. The DASH (Disability of Arm-Shoulder-Hand) outcome]. Handchir Mikrochir Plast Chir, 31(3):149-152.

WARE JE,Jr, Sherbourne CD. 1992. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care, (6):473-483.

Part C

MANUSCRIPT (PUPLICATION-READY FORMAT)

ORIGINAL ARTICLE

Reliability of shoulder symptom recall after 1 year in a retrospective application of the Oxford Shoulder Score

Michael Held, Steve Roche, Basil Vrettos, Maritz Laubscher & Johan Walters
Department of Orthopaedic Surgery, Groote Schuur Hospital, University of Cape Town, Cape Town, South Africa

Received
Received 21 June 2011;
accepted 9 February 2012

Keywords
Oxford Shoulder Score, recall, retrospective, reliability, patient-based outcome scores, measures of recovery.

Correspondence
Michael Held,
Department of Orthopaedic,
Groote Schuur Hospital,
University of Cape Town,
H49 Old Main Building,
Cape Town 7925, South Africa.
Tel.: +27 214066157
Fax: +27 21472709.
DOI:10.1111/j.1758-5740.2012.00189.x

ABSTRACT

Background The accuracy of retrospective recall of shoulder symptoms has not been well documented. This prospective study assesses the ability of patients to recall their preoperative shoulder function one year after a surgical intervention, using the Oxford Shoulder Score (OSS).

Methods 35 patients completed an OSS before undergoing shoulder surgery. One year later, patients were asked to recall their symptoms prior to their surgery. The recalled OSS of the patients as a group was compared to their preoperative score. The recall bias of each test pair was assessed with a Bland – Altman plot.

Results On recall after a mean of 12.6 months, the mean OSS from the index assessment increased from 36.25 to 38.25 points. The mean difference of 2 points for the patients as a group was not significant ($p = 0.14$). The statistical limits of agreement of the Bland – Altman plot were set at $+/-2$ SD = 14.079 points. The plotted points showed fair correlation between each individual test pair.

Conclusion The recall of symptoms of a large group of patients at 1 year after the index intervention appears to have a moderate correlation with the preoperative scoring. Although statistically acceptable, this limit of agreement is much larger than the 4.5-point difference, established to be clinically relevant in prior studies. The variation seen within the scores at the individual level suggests that these data cannot be used as a retrospective tool.

INTRODUCTION

The accurate assessment of the functional status and symptoms experienced by patients is essential for clinical evaluation before and after surgery. Quantitative scoring systems, as well as patient-reported outcome measures such as the Oxford Shoulder Score (OSS), comprise tools for enhancing this process [1]. Unfortunately, we are often faced with the challenge that the pre-operative documentation of the shoulder function of our patients is inadequate. This problem is fuelled by the large number of trauma patients in whom prospective scoring is impossible, as well as poor documentation from medical staff and problems of data retrieval from dysfunctional filing systems. Despite these circumstances, we would be able to compare pre-operative and postoperative shoulder function in patients if they could remember their symptoms before their surgery, even at a later stage of their postoperative evaluation. A study by Wilson et al. reported an overall significant correlation between pre- and postoperative assessments of a cohort of patients conducted within 2 months [2]. However, wide variation was recorded in the assessment of individuals. In the present study, we aimed to investigate the reliability of the recall of symptoms using the OSS when applied retrospectively, at a mean of 12 months after surgery.

MATERIALS AND METHODS

The present prospective study included 35 adult patients who were under the care of one dedicated shoulder surgeon. Each patient completed the OSS, 1 day before the operation. Ethics approval was obtained from the institutions Ethics Committee. One year after surgery, patients were contacted by telephone and were asked to recall their symptoms during the days before their shoulder surgery. Patients aged < 18 years and patients with cognitive impairment were excluded from enrolment. All patients gave their consent before participation. The procedures performed included arthroscopic subacromial decompression, acromio-clavicular joint excision, rotator cuff repair, surgical treatment of calcific tendinitis and total joint replacement of the shoulder.

The OSS is a 12-item scoring tool designed for patients undergoing shoulder operations (excluding shoulder stabilization). Each item is graded from 1 to 5, adding to give a single total score. The best score is 12 and the worst is 60. Pain is assessed through four items. Activities of daily living are assessed through the remaining eight items. The score has been validated and is sensitive to change with good test–retest reliability at 24 hours [3].

The Shapiro–Wilk test was used to determine whether the data were distributed normally or not. A paired *t*-test was used to compare the pre-operative and recalled OSS. A Bland–Altman plot was used to assess the repeatability of the OSS at two time points and to demonstrate that variability is not related to the magnitude of the score. Statistical analysis was performed using STATA, version 12.0 (StataCorp, College Station, TX, USA).

RESULTS

Thirty-five patients were recruited into the present study; however, seven patients (20%) could not be traced to assess the

recalled OSS. The untraceable patients were not significantly different from those who could be contacted with respect to age and gender distribution; the mean (SD) age of the patients who could be contacted was 61 (8) years (range 43 years to 75 years) compared to patients lost to follow-up, whose mean (SD) age was 59 (13) years (range 41 years to 78 years) (two-sided t-test, $p = 0.6, t = 0.5$), and the proportion of men among the followed up patients was 68% [95% confidence interval (CI) = 48% to 84%] compared to 71% (95% CI = 29% to 96%, $p = 0.9$). Taking the small number of patients into account, the mean pre-operative OSS score for the untraceable patients was lower than for the traceable patients with a mean (SD) of 29 (9.6) points compared to 36.25 (7.6) points; using a two-sided t-test, this was of borderline significance ($t = 2.03, p = 0.0502$). The remainder of the results are presented for the 28 patients who were followed up.

At a mean of 12.6 months (range 10 months to 18 months), patients who could be contacted recorded their mean (SD) pre-operative shoulder function score as 38.25 (7.4) points (range 25 to 57). The mean difference of two points (95% CI = −0.73 to 4.73) ranged from −10 to 13 (SD = 7.0) and was not statistically significant ($p = 0.14$). Figure 1 shows a linear relationship, with fair correlation, on comparing the initial test scores with the recollected scores as a collective (Pearsons rho 0.560, $p < 0.001$).

A Bland–Altman plot [4] shows bias when comparing the two tests with each other and was used to assess the test pair of each patient (Fig. 2). The limits of agreement were set at two SDs from −12.079 to 16.079, including all outliers. The difference did not vary according to the value of the mean and a Pitman test showed no significant linear correlation between the difference and the mean ($r = -0.016$, $n = 28$, $p = 0.936$).

DISCUSSION

The ability for retrospective recall of symptoms has significant impact in situations where this information cannot be collected before surgical intervention or, indeed, if the patient is newly acquainted with the treating physician. The outcome of any intervention is affected to a greater or lesser extent by the pre-existing clinical condition. The present study aimed to assess the ability of a cohort of patients undergoing shoulder surgery to recall their pre-operative status 1 year after the index operation. To avoid a heightened consciousness, the patients included in the present study were unaware that they would be asked to recall the symptoms that they had experienced before the operation. The lower pre-operative OSS of the patients lost to follow-up (mean of 29 compared to 36.25) is unlikely to bias the results. Figures 1 and 2 show that the recall of patients who could be contacted was not influenced by a lower or higher pre-operative OSS. A comparison of the initial test scores and the recall scores at 1 year as a collective showed fair correlation, with a Pearson's rho of 0.56 (Fig. 1). Similarly, the individual test agreement was moderate, with an intraclass correlation of 0.54. Bland and Altman developed a plot to show bias when comparing two different tests with each other [4]. The difference of each test pair is plotted against their mean. The 0-line indicates a difference of 0 points; hence, two pairs produce equal results. Statistically, the limits of agreement are defined as the mean difference ± 2 SD of the differences, which indicates that 95% of the values are normally distributed. In our Bland–Altman plot, the mean difference of the individual test pairs was two points and the limit of ± 2 SD was 14.079 points or −12.079 and 16.079 points, respectively, if added or subtracted from the mean difference (two points). In Fig. 2, the plotted points, which could be accommodated within these limits, show fair correlation when comparing the initial OSS with the recalled OSS of each test pair.

In a previous study by Wilson et al., the reliability of shoulder symptom recall was evaluated and it was concluded to be valid for assessing the impact of a shoulder operation in a large group of patients within a population [2] (Fig. 3). The time to recall after the

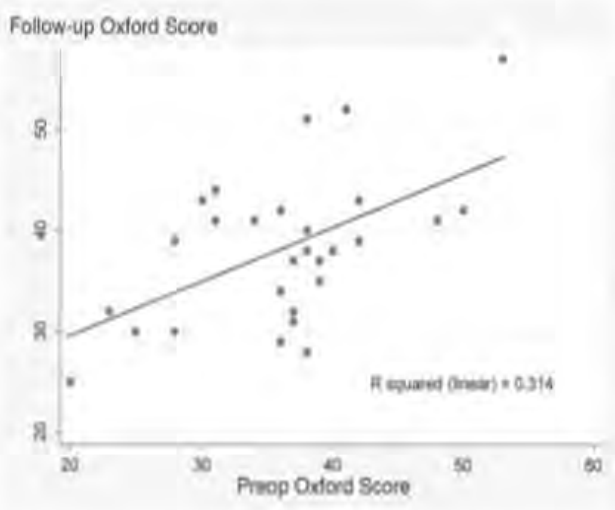

Fig. 1 Scatter plot, indicating fair correlation between the pre-operative and the follow-up Oxford Shoulder score.

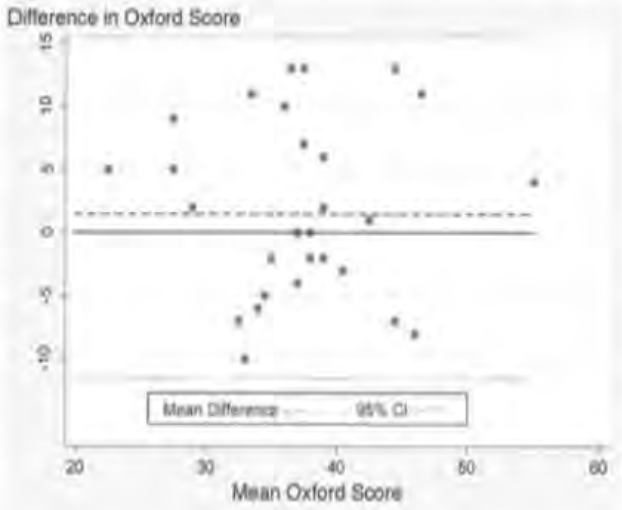

Fig. 2 Bland–Altman plot, showing the difference of each individual test pair against their mean Oxford Shoulder score. The 0-line indicates a difference of 0 points; hence, each test pair produces equal results. The statistical limits of agreement (± 2 SD = 14.079 points) is added or subtracted from the mean difference (two points). The plotted points, which can be accommodated within these limits, indicate fair correlation when comparing the initial OSS with the recalled OSS of each test pair.

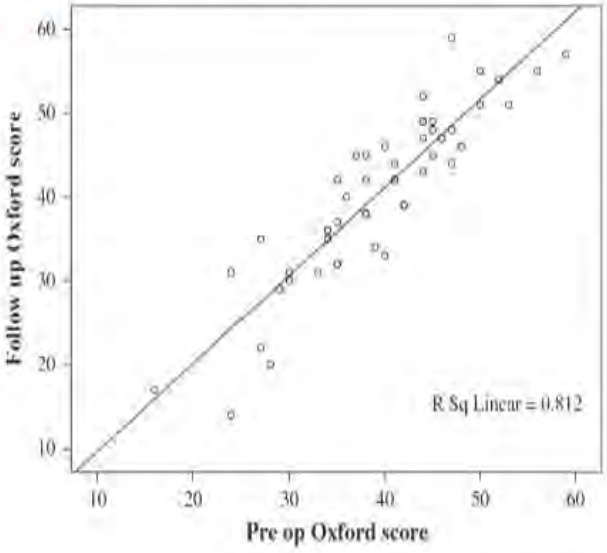

Fig. 3 Scatter plot of the study of Wilson et al., showing the correlation of the pre-operative Oxford Shoulder score compared to the score on follow-up, when remembering previous symptoms [2].

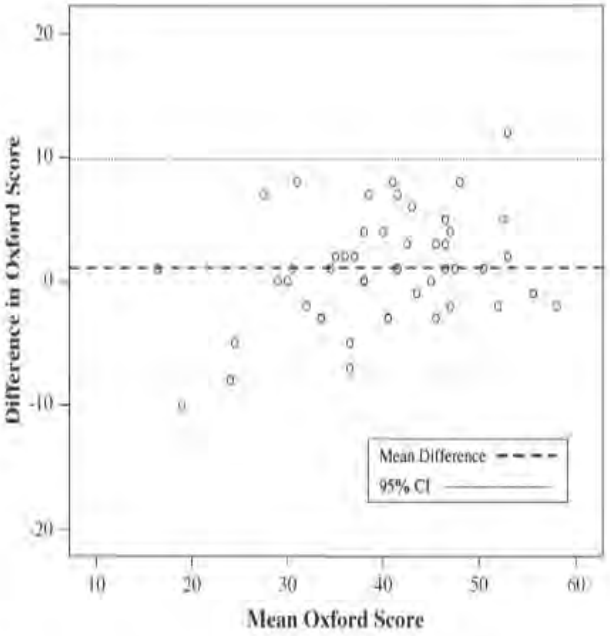

Fig. 4 Bland–Altman plot of the study of Wilson et al., showing the difference of each individual test pair plotted against the mean Oxford Shoulder score [2].

surgery was much shorter in the study by Wilson et al. [2] (50 days, range 22 days to 150 days) compared to that of the present study (12.6 months or 378 days; range 299 days to 541 days). The mean age of the 63 patients included was 57 years (range 20 years to 86 years), which is similar to the present study (mean age of 61 years). Wilson et al. also compared patients aged < 50 years with patients aged >60 years and found that age did not adversely influence the recall [2].

However, based on the BA plot (Fig. 4), poor correlation was found when comparing the individual scores of each patient. Outliers exceeded the limits of agreement that Wilson et al. set at a 4.5 point difference, which is considered clinically relevant by Cloke et al. [5]. Adapting these limits of agreement, we are also concerned whether the discrepancy of 2 SD is appropriate in the present study. This would mean that a positive correlation between the two test scores (i.e. the initial score and the recalled score) is a random occurrence.

The responses of each individual question were also compared with each other to evaluate whether particular questions were recalled better than others. However, these results should be interpreted with caution given the relatively small data set. Applying the Shapiro–Wilk test, a normal distribution of data was found in eight of the 12 questions; however, *p*-values were estimated using the nonparametric Wilcoxon test because each question employed a five-point scale. Using the *t*-test lead to the same conclusions as the nonparametric test.

When comparing the recalled scores with pre-operative scores of each question, statistically significant differences were found in the answers to questions 4, 7, 8, 9 and 12 (using knife and fork, combing hair, baseline pain, hanging up clothes, pain at night), indicating that the patients could not recall their answers to these questions. Questions 1, 2, 3, 5, 6, 10 and 11 (maximum pain, dressing, using private or public transport, household shopping, carrying a plate of food, washing, working) led to answers that were not statistically significantly different when asked 12 months later. This demonstrates a trend, indicating that limitations in simple daily tasks, pain at night and general baseline pain are not recalled reliably.

The present study has some weaknesses. Similar to the study by Wilson et al., only one assessment tool, namely the OSS, was used [2]. Other scoring tools offering more variety or a more general assessment (Disabilities of the Arm, Shoulder and Hand; 36-Item Short Form Health Survey) [6,7]) may provide a benefit. In the present study, the interval between the two tests was 1 year compared to 6 weeks in the study by Wilson et al. thereby eliminating one found in the previous study [2]. The longer time period in the present study appears to be more appropriate because it allowed for complete healing and rehabilitation after the operation. Thus, patients would have to recall their symptoms, rather than being reminded by residual pain or functional deficits.

Conclusions

From the data obtained in the present study, it is concluded that the recall of symptoms of a large group of patients at 1 year after the index intervention appears to have a moderate correlation with that recorded immediately before the operation. Adopting clinically relevant limits of agreement, the variation seen within the scores of each individual suggests that overall outcome may be a random finding, and that patient recall of symptoms at 1 year cannot be used as a retrospective tool. Knowing that our patients cannot recall the severity of their pre-operative shoulder symptoms reliably, we might need to reassess how much relevance we place on comparing them with their postoperative scores. In future studies, we should evaluate at what point in time after the operation a patient can no longer recall his pre-operative shoulder symptoms because this might represent the point when, subjectively, recovery takes place. It is possible that these findings

will be applicable to clinical areas beyond the field of orthopaedic surgery, which rely on patient-based clinical outcome measures.

ACKNOWLEDGEMENTS

We acknowledge the support of Henri Carrara for his assistance with the analysis in the present study.

References

1. **Dawson J, Rogers K, Fitzpatrick R, Carr A.** The Oxford shoulder score revisited. *Arch Orthop Trauma Surg* 2009; 129:119–23.
2. **Wilson J, Baker P, Rangan A.** Is retrospective application of the Oxford Shoulder score valid? *J Shoulder Elbow Surg* 2009; 18: 577–80.
3. **Dawson J, Fitzpatrick R, Carr A.** Questionnaire on the perceptions of patients about shoulder surgery. *J Bone Joint Surg Br* 1996; 78:593–600.
4. **Bland JM, Altman DG.** Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1:307–10.
5. **Cloke DJ, Watson H, Purdy S, Steen IN, Williams JR.** A pilot randomized, controlled trial of treatment for painful arc of the shoulder. *J Shoulder Elbow Surg* 2008; 17(Suppl 1):17S–21S.
6. **Germann G, Wind G, Harth A.** The DASH (Disability of Arm-Shoulder-Hand) Questionnaire – a new instrument for evaluating upper extremity treatment outcome. *Handchir Mikrochir Plast Chir* 1999; 31:149–52.
7. **Ware JE Jr, Sherbourne CD.** The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; 30:473–83.

Part D

APPENDICES

I. Acknowledgements

Description of roles played by each co-author:

J. Walters:	supervisor and editor of final script
S. Roche:	supervisor
B. Vrettos:	provided database of patients
M. Laubscher:	assisted with data collection of recall shoulder scores

II. Ethics Approval

UNIVERSITY OF CAPE TOWN

Faculty of Health Sciences
Human Research Ethics Committee
Room E52-24 Groote Schuur Hospital Old Main Building
Observatory 7925
Telephone [021] 406 6626 • **Facsimile** [021] 406 6411
e-mail: lamees.emjedi@uct.ac.za

09 June 2010

HREC REF: 216/2010

Dr M Held
C/o Dr S Roche
Orthopaedics
GSH

Dear Dr Held

PROJECT TITLE: Reliability of shoulder symptom recall using a retrospective application of the Oxford Shoulder Score
Supervisor: S Roche (Orthopaedics)

Thank you submitting your study to the Faculty of Health Sciences Human Research Ethics Committee.

It is a pleasure to inform you that the Ethics Committee has **formally approved** the above-mentioned study.

Approval is granted until 15 July 2012.

Please send us an annual progress report (website form FHS 016: http://www.health.uct.ac.za/research/humanethics/forms/) if your research continues beyond the approval period. Alternatively, please send us a brief summary of your findings so that we can close the research file.

Please note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Please quote the REC. REF in all your correspondence.

Yours sincerely

Signed by candidate

PROFESSOR MARC BLOCKMAN
CHAIRPERSON, FHS human research ethics committee

Federal Wide Assurance Number: FWA00001637.
Institutional Review Board (IRB) number: IRB00001938

lemjedi

III. Consent Form

The patient will receive the following information prior to the publication of his case report, if he/she agrees.

DESCRIPTION: We would like to publish the course and management of your case as your injury is very rare and will give others valuable information on how to treat similar injuries in the future.
RISKS AND BENEFITS: there are no anticipated risks associated with this study. You will not receive any direct benefit from participation.
TIME INVOLVEMENT: Your participation in this study will not require 15 minutes.
PAYMENTS: You will not be paid to participate in this study.
PARTICIPANT'S RIGHTS: Your decision whether or not to participate in this study will not affect your medical care. Your participation is voluntary and you have the right to withdraw your consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. Your identity will not be disclosed in any published and written material resulting from the study.

Who may use or disclose the information?

The following parties are authorized to use and/or disclose your health information in connection with this research study:

1. The Protocol Director of the University of Cape Town Administrative Panel on Human Subjects in Medical Research.
2. Research Staff working on this project.

Your personal information will be deleted from the data stream.

Contact Information:
Questions, Concerns, or Complaints: If you have any questions, concerns or complaints about this research study, its procedures, risks and benefits, or alternative courses of treatment, you should ask the Chairperson of the Human Research Ethics Committee, Prof Blockman: 021 406 6492.

IV. Questionnaire (Oxford Shoulder Score)

PROBLEMS WITH YOUR SHOULDER

During the past 4 weeks.....

✓tick one box for each question

1. *During the past 4 weeks.....*
How would you describe the *worst* pain you had from your shoulder?

None	Mild	Moderate	Severe	Unbearable
❑	❑	❑	❑	❑

2. *During the past 4 weeks.....*
Have you had any trouble dressing yourself because of your shoulder?

No trouble at all	A little bit of trouble	Moderate trouble	Extreme difficulty	Impossible to do
❑	❑	❑	❑	❑

3. *During the past 4 weeks.....*
Have you had any trouble getting in and out of a car or using public transport because of your shoulder?

No trouble at all	A little bit of trouble	Moderate trouble	Extreme difficulty	Impossible to do
❑	❑	❑	❑	❑

4. *During the past 4 weeks.....*
Have you been able to use a knife and fork - at the same time?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
❑	❑	❑	❑	❑

5. *During the past 4 weeks.....*
Could you do the household shopping on your own?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
❑	❑	❑	❑	❑

6. *During the past 4 weeks.....*
Could you carry a tray containing a plate of food across a room?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
❑	❑	❑	❑	❑

During the past 4 weeks.....

✓tick **one** box for each question

7. *During the past 4 weeks.....*

Could you brush/comb your hair with the affected arm?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
❑	❑	❑	❑	❑

8. *During the past 4 weeks.....*

How would you describe the pain you usually had from your shoulder?

None	Very mild	Mild	Moderate	Severe
❑	❑	❑	❑	❑

9. *During the past 4 weeks.....*

Could you hang your clothes up in a wardrobe, - using the affected arm?

Yes, Easily	With little difficulty	With moderate difficulty	With great difficulty	No, Impossible
❑	❑	❑	❑	❑

10 *During the past 4 weeks.....*

Have you been able to wash and dry yourself under both arms?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
❑	❑	❑	❑	❑

11 *During the past 4 weeks.....*

How much has pain from your shoulder interfered with your usual work *(including housework)*?

Not at all	A little bit	Moderately	Greatly	Totally
❑	❑	❑	❑	❑

12 *During the past 4 weeks.....*

Have you been troubled by pain from your shoulder in bed at night?

No nights	Only 1 or 2 nights	Some nights	Most nights	Every night
❑	❑	❑	❑	❑