Arch Orthop Trauma Surg (2009) 129:119–123
DOI 10.1007/s00402-007-0549-7

ORTHOPAEDIC OUTCOME ASSESSMENT

The Oxford shoulder score revisited

Jill Dawson · Katherine Rogers · Ray Fitzpatrick · Andrew Carr

Received: 11 October 2007 / Published online: 9 January 2008

Abstract The validated, patient-reported Oxford shoulder score (OSS) was introduced around 10 years ago, primarily for the assessment of outcomes of shoulder surgery (excluding shoulder stabilisation) in randomised trials. Its uptake has steadily increased in a number of countries and its use has also been extended. Recently a number of issues have been raised in relation to other related patient-reported outcome measures which were devised around the same time as the OSS. This included recommendations to change the scoring system. This paper reviews issues concerning patient-reported outcome measures that apply to the OSS and makes some recommendations (including changes to the scoring system) as to how it should be used.

Keywords Shoulder · Outcome score · Shoulder surgery · Oxford shoulder score · Patient-reported outcomes

Introduction

It is now around 10 years since the patient-reported Oxford shoulder score (OSS) was introduced [1]. At the time of its development, while a number of clinician-devised or assessed scores existed for the shoulder (e.g. the Constant–Murley shoulder score [2], Simple Shoulder Test [3]) the use of patient-reported measures, which represented the patient's perspective, in orthopaedics, was extremely limited. Devised with patients, the OSS was primarily developed for the assessment of outcomes of shoulder surgery (excluding shoulder stabilisation, for which there is a separate, specific patient-reported measure—the Oxford Shoulder Instability Score [4]) in randomised trials and was designed to be completed by the patient, in order to minimise potential reporting bias (e.g. bias unwittingly introduced by surgeons assessing their own patients' outcomes [5]). An additional advantage in using patient-reported outcome measures (PROMs) is that, unlike a clinical assessment, they can be completed at a remote location by post, thereby avoiding inconvenience and cost to all concerned.

The OSS was devised as a joint specific instrument so as to minimise the influence of other co-morbidity and underwent rigorous assessment of reliability, validity and responsiveness within prospective studies [1, 6]. Over the years its uptake has steadily increased and it has now been used in a number of countries (including the UK [7], Israel [8], Finland [9]). Details of a study to formally translate and validate the OSS in German have also been published [10]. Recently, the OSS has been adopted as the primary outcome measure in a UK RCT of interventions for rotator cuff tears [11]. The use of the OSS has also been extended and it has been applied in cohort studies, audits and a national joint replacement registry [12].

Recently a number of issues have been raised in relation to other patient-reported outcome measures (the Oxford hip and knee scores [13–15)] which were devised (involving members of our team) at around the same time as the OSS and which have a similar scoring system. One issue included the suggestion that the scoring system needs to change, which has had the potential to cause some confusion. The aim, therefore, of this paper is to review the situation with the OSS and make some recommendations as to how it should be used.

J. Dawson (✉) · K. Rogers · R. Fitzpatrick
Department of Public Health, University of Oxford,
Old Road Campus, Oxford OX37LF, UK
e-mail: jill.dawson@dphpc.ox.ac.uk

A. Carr
Nuffield Department of Orthopaedic Surgery,
University of Oxford, Windmill Road,
Oxford OX37LD, UK

Wording of questions and response categories

The recommended format for OSS questions and their response categories can be found on the Patient-Reported Health Instruments web-site [16]. We have found that response rates for the OSS are generally high (this may partially reflect the average—generally middle—age of patients undergoing shoulder surgery). For example, 96% of patients completed all OSS items at both pre- and 6 month post-operative assessments in one study [17]. However, Occasionally, patients have difficulties answering particular questions. For instance the question (5) "Could you do the household shopping on your own?" may be difficult to answer for people who say that someone else does their shopping or for people living in residential care. The word "could" suggests that patients should answer this item hypothetically in such circumstances and this is our recommendation, which can be communicated to the patient. A similar approach should be taken for the question about brushing/combing hair, by patients who are bald headed. If, after clarification, an item is simply left unanswered, then this should be dealt with as missing data (discussed below).

The questions in the OSS represent issues found to be of general importance in the initial interviews conducted with patients, from which the OSS resulted [1], and which were relevant to the vast majority of patients—which they need to be when included in an outcome measure. However, it is wrong to imagine that a perfect questionnaire, which suits all of the people, all of the time, will ever exist.

Scoring

When the OSS was originally devised, the scoring system was designed to be as simple as possible, in order to encourage its use. Thus, each of the 12 questions was scored from 1 to 5, with 1 representing best outcome/least symptoms. Scores from each question were added so the overall score was from 12 to 60 with 12 being the best outcome. This was identical to the system used for the Oxford hip (OHS) and knee scores (OKS). However, subsequently, many surgeons using the OHS or OKS said that they found this scoring unintuitive and started using different systems which led to some confusion. We therefore issued recent recommendations concerning changes to the method of scoring the Oxford hip and knee scores [15]. Our view is that similar changes are inevitable with the OSS and that therefore, the sooner this change occurs, the better. Under the new system, each question on the OSS should be scored 0–4, with four representing the best (this is the opposite direction from the original method of scoring). When the 12 items are summed, this produces overall scores that run from 0 to 48 with 48 being the best outcome (to convert the "old system" of 60–12 to the 0–48 scoring system and vice versa simply subtract the score from 60) [18]. In addition, the method used should always be clearly stated (including in abstracts). We also recommend that this scoring system shoulder be adopted for the Oxford Shoulder Instability Score [4].

Comparison with other scoring systems

The OSS item content developed out of interviews conducted with patients and therefore fully reflects the patient's perspective. As far as we can ascertain, while the Western Ontario Rotator Cuff Index (WORC) [19] involved some interviews with patients during the development stage, no other patient-reported outcome measures for the shoulder have [3, 20–25]. Giving prominence to the patient's—rather than the clinician's—perspective is preferable because patients and clinicians can genuinely disagree about the relative importance of different aspects of the outcome of health care interventions [26, 27], and because the patient's perspective on outcomes might be expected to more closely mirror their attitudes regarding satisfaction with their treatment.

The OSS was designed to be joint specific in order that it should be as sensitive to the outcome of shoulder surgery as possible and to be influenced as little as possible by other co-morbidities ("noise")—although it is impossible to totally eradicate the affect of noise on even quite specific measures of outcomes. This feature of a score, its specificity, influences its responsiveness or "sensitivity to change", which is the most important aspect in relation to prospective outcome studies [28, 29]. The OSS has been shown to have particularly high responsiveness [1, 6, 20] that is comparable to the clinician assessed Constant–Murley score [6].[1] It likely also has measurement properties that are generally superior to older patient-reported instruments which were developed at a time when psychometric methods to develop and test new instruments were not well appreciated or applied [20]. Due to the fact that the OSS has been evaluated independently and found to be a highly reliable and responsive system for the assessment of shoulder surgery [7, 20], there is some justification for using this score in isolation. However, if it is important to compare the improvement resulting from shoulder interventions with interventions occurring at other sites, it is sensible to use a general health measure, such as the SF-36

[1] Although one series of analyses revealed patient-reported outcome measures (OSS and SF-36) to be more stable than the Constant–Murley assessment, based on comparisons with ratings on satisfaction and transition items [6].

Table 1 Mean Oxford shoulder scores before and at 12 months after (NHS funded) shoulder surgery, by 10% bands (deciles) of pre-operative score

Pre-operation Oxford shoulder score band	12–60 scale	0–48 scale	Mean (SD) Oxford Shoulder Score (Using 0–48 scoring method)				Change in score between 0 and 12 months
			Before operation		12 months after		
1	12–23	37–48 Best possible	39.4 (1.7)	$n = 14$	43.6 (4.8)	$n = 12$	4.0 (4.6)
2	24–26	34–36	34.9 (0.9)	$n = 17$	40.5 (4.2)	$n = 15$	5.6 (4.3)
3	27–29	31–33	32.4 (0.8)	$n = 15$	38.6 (8.3)	$n = 12$	6.0 (8.1)
4	30–32	28–30	29.2 (0.8)	$n = 24$	36.7 (8.4)	$n = 19$	7.4 (8.6)
5	33–34	26–27	26.6 (0.5)	$n = 19$	40.8 (6.3)	$n = 18$	14.2 (6.2)
6	35–36	24–25	24.5 (0.5)	$n = 15$	32.7 (8.0)	$n = 11$	8.3 (7.9)
7	37–39	21–23	21.9 (0.9)	$n = 14$	39.8 (5.8)	$n = 11$	17.8 (6.0)
8	40–43	17–20	18.7 (1.1)	$n = 18$	29.9 (11.1)	$n = 15$	11.2 (11.1)
9	44–48	12–16	14.2 (1.6)	$n = 17$	23.6 (9.9)	$n = 16$	9.4 (9.7)
10	49–60	0–11 Most severe	8.5 (2.5)	$n = 17$	23.4 (11.9)	$n = 14$	14.6 (12.1)
Total			24.9 (9.0)	$n = 170$	34.8 (10.6)	$n = 143$	9.9 (9.0)

[30, 31], as well as the OSS. If health economic information is needed the EuroQol [32] is valuable. In addition if specific clinical and surgical data, such as range of movement, is required then a formal clinical assessment would also be necessary.

Use of the score

Although the OSS was designed to be a primary outcome measure in randomised controlled trials, it has also been used in cohort studies and audits. It has become apparent that one of the biggest determinants of outcome after some forms of orthopaedic surgery is the pre-operative score [15, 33, 34], and this finding may also hold true for shoulder surgery. Therefore if the treatment of different cohorts of patients is being studied, in a non-randomised setting, it is essential that both the preoperative and postoperative scores are obtained. The change in the score should be analysed in addition to the post-operative score. Likewise, if a multivariate analysis of outcomes is undertaken this should take account of the pre-operative score [35]. Following shoulder surgery, the majority of the improvement in pain and function and in the OSS likely occurs within the first year [17]. It is therefore not unreasonable to assess outcome at one year.

Given that the 1-year score is related to the preoperative score; when using the scores for audit purposes, it is useful to know approximately what outcome would be expected after shoulder surgery, given a particular preoperative score. Results from an analysis of data taken from the original study that devised the OSS [1], are shown in Table 1. This summarises mean OSS before and after shoulder surgery, presented by 10% bands (deciles) of the pre-operative score.[2] These data are also shown converted to the 0–48 scoring system. The table shows that those patients who started with a worse OSS before surgery tended to remain worse after the operation. It should, however, be noted that with all outcome scores, scores tend to worsen with age [36, 37]. Therefore, for each condition or type of surgery that is studied, in elderly patients, a "normal" score may be somewhat less than 48.

Although we have achieved very high response rates when the OSS has been sent to patients, practical approaches to maximise response rates should always be employed. These include using carefully worded covering letters, sending reminders with second copies of the questionnaire, using prepaid reply envelopes and contacting patients by telephone. When patients have bilateral joint problems we favour giving two questionnaires, one for each side, rather than, for instance, modifying the questionnaires to include both sides.

Surgeons are attracted to categorisation systems, grouping patients' results according to whether they are considered excellent, good, fair or poor, rather than simply using a score. However, such categories (or rather, their cut-off points) are, unfortunately, always very approximate and are likely to vary from one population to another. Work is currently in progress to produce categories. Until this is available we believe that surgeons should avoid categorisation.

Other languages

The OSS is now being used in other countries with details of formal translation and validation procedures published,

[2] This secondary data analysis was conducted on anonymous data which had been retained from the original study [1].

so far, for a German translation study [10]. If data from the OSS are to be used comparatively across different languages then the translational process has to be performed in an appropriate, standard manner as described by Huber et al. [10]. This should include forward and backward translation methods, plus an assessment of the translated score's measurement properties.

Missing data

A common problem with questionnaires is that some patients provide incomplete responses. We propose that, if, after repeated attempts to obtain complete data from an individual, only one or two questions have been left unanswered, it is reasonable to enter the mean value representing all of their other responses, to fill the gaps. An alternative computerised method of imputing values, which could be applied to many questionnaires has been reported by Jenkinson et al. [38]. If more than two questions are unanswered we believe that an overall score should not be calculated. If patients indicate two answers for one question we recommend that the convention of using the worst (most severe) response is adopted.

Statistical issues

By the time patients are close to receiving shoulder surgery, their symptoms will often be quite severe, whereas, by 12 months following surgery the majority of patients generally have only very mild problems—if any. For these reasons, data from the OSS obtained at these time-points are generally skewed in one or other direction [1]. It could therefore be argued that it is sensible to use transformations [39] or non-parametric statistics for analyses involving absolute scores. Analysis using change scores is less problematic as they tend to be more normally distributed.

While it is simple to determine the statistical significance of changes in health status measures—such as the OSS, it can be harder to determine the real clinical or subjective meaning of these changes. There are a number of approaches to determining the smallest amount of change on a measure that is likely to be of importance [40, 41], which include the minimal clinically important difference (MCID) [40]. A MCID is the smallest change in score which patients perceive as meaningful and which would cause clinicians to consider a change in the patient's management [42]. Work is in process to produce MCID estimates for the OSS. Until these are available an approximation to the MCID can be obtained, based on the observation that for many PROMs the MCID is about half of the standard deviation of change [43].

One of the reasons why clinically important differences are important is that this information is needed in power calculations to determine the size of a study. It is essential that power calculations are done before a study is undertaken.

As patients' involvement in outcome assessment is becoming more widely established, it is becoming increasingly important to achieve some standardisation in the use of instruments. We trust that this paper will assist in that regard.

Acknowledgments Regarding the secondary data analysis conducted to inform the table in this paper. The original study that generated these data was conducted in the early 1990s and we wish to acknowledge receipt of funding for that study by grant from Oxford Regional Health Authority (Audit). These data were retained in an anonymised form. The study complied with the laws of the UK, which at that time, did not require informed consent from patients for a purely observational study, as completion of a questionnaire was accepted as implicit consent.

Conflict of interest statement None of the authors have any conflict of interest in relation to this paper.

References

1. Dawson J, Fitzpatrick R, Carr A (1996) Questionnaire on the perceptions of patients about shoulder surgery. J Bone Joint Surg (Br) 78:593–600
2. Constant CR, Murley AH (1987) A clinical method of functional assessment of the shoulder. Clin Orthop (214):160–164
3. Lippitt SB, Harryman DT, Matsen FA (1993) A practical tool for evaluating function: the simple shoulder test. The shoulder: a balance of mobility and stability. American Academy of Orthopaedic Surgeons, Rosemont, pp 501–518
4. Dawson J, Fitzpatrick R, Carr A (1999) The assessment of shoulder instability: the development and validation of a questionnaire. J Bone Joint Surg (Br) 81-B:420–426
5. Pynsent P, Fairbank JTC, Carr A (1993) Outcome measures in orthopaedics. 1st edn. Butterworth-Heinemann, Oxford
6. Dawson J, Hill G, Fitzpatrick R, Carr A (2002) Comparison of clinical and patient-based measures to assess medium-term outcomes following shoulder surgery for disorders of the rotator cuff. Arthritis Rheum 47(5):513–9
7. Cloke DJ, Lynn SE, Watson H, Steen IN, Purdy S, Williams JR (2005) A comparison of functional, patient-based scores in subacromial impingement. J Shoulder Elbow Surg 14(4):380–384
8. Rosenberg N, Soudry M (2006) Shoulder impairment following treatment of diaphysial fractures of humerus by functional brace. Arch Orthop Trauma Surg 126:437–440
9. Flinkkila T, Ristiniemi J, Lakovaara M, Hyvonen P, Leppilahti J (2006) Hook-plate fixation of unstable lateral clavicle fractures. Acta Orth 77(4):644–649
10. Huber W, Hofstaetter JG, Hanslik-Schnabel B, Posch M, Wurnig C (2004) The German version of the Oxford shoulder score—cross-cultural adaptation and validation. Arch Orthop Trauma Surg 124(8):531–536
11. Carr AJ, Fitzpatrick R, Gray A, Dawson J, Norrie J, Campbell M, Ramsay C, Rees J, Moser J (2007) United Kingdom Rotator Cuff Study (UKUFF trial) 01/05/2007–30/04/2012. Web-site: https://viis.abdn.ac.uk/HSRU/UKUFF/Site/Public/Default.aspx. HTA reference 05/47/02. Funded by the Department of Health

12. New Zealand National Joint Registry (2007) Canterbury District Health Board, New Zealand. http://www.cdhb.govt.nz/NJR/
13. Dawson J, Fitzpatrick R, Carr A, Murray D (1996) Questionnaire on the perceptions of patients about total hip replacement. J Bone Joint Surg (Br) 78:185–190
14. Dawson J, Fitzpatrick R, Murray D, Carr A (1998) Questionnaire on the perceptions of patients about total knee replacement. J Bone Joint Surg (Br) 80(1):63–69
15. Murray DW, Fitzpatrick R, Rogers K, Pandit H, Beard DJ, Carr AJ, Dawson J (2007) The use of the Oxford hip and knee scores. J Bone Joint Surg (Br) 89-B:1010–1014
16. Unit of Health-Care Epidemiology (2007) See Oxford Orthopaedic scores link on patient-reported health instruments. University of Oxford, web-site http://phi.uhce.ox.ac.uk/
17. Dawson J, Hill G, Fitzpatrick R, Carr A (2001) The benefits of using patient-based methods of assessment: medium term results of an observational study of shoulder surgery. J Bone Joint Surg (Br) 83(6):877–882
18. Weale AE, Halabi OA, Jones PW, While SH (2001) Perceptions of outcomes after unicompartmental and total knee replacements. Clin Orthop Rel Res 382:143–153
19. Kirkley A, Griffin S, Alvarez C (2003) The development and evaluation of a disease-specific quality of life measurement tool for rotator cuff disease: the Western Ontario rotator cuff index (WORC). Clin J Sport Med 13:84–92
20. Kirkley A, Griffin S, Dainty K (2003) Shoulder systems for the functional assessment of the shoulder. Arthroscopy 19(10):1109–1120
21. Amstutz HC, Sew Hoy AL, Clarke IC (1981) UCLA anatomic total shoulder arthroplasty. Clin Orthop 155:7–20
22. Roach KE, Budiman ME, Songsiridej N, Lertratanakul Y (1991) Development of a shoulder pain and disability index. Arthritis Care Res 4(4):143–149
23. Hudak PL, Amadio PC, Bombardier C (1996) Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) (corrected). The Upper Extremity Collaborative Group (UECG) [Published erratum appears in Am J Ind Med (1996) 30(3):372]. Am J Ind Med 29(6):602–608
24. L'Insalata JC, Warren RF, Cohen SB, Altchek DW, Peterson MG (1997) A self-administered questionnaire for assessment of symptoms and function of the shoulder. J Bone Joint Surg (Am) 79(5):738–748
25. Boehm D, Wollmerstedt N, Doesch M, Handwerker M, Mehling E, Gohlke F (2004) Development of a questionnaire based on the Constant–Murley score for self-evaluation of shoulder function by patients. Unfallchirurg 107(5):397–402
26. Wright JG, Rudicel S, Feinstein AR (1994) Ask patients what they want. Evaluation of individual complaints before total hip replacement. J Bone Joint Surg (Br) 76(2):229–234
27. Schneider W, Knahr K (2001) Surgery for Hallux Valgus. The expectations of patients and surgeons. Int Orthop 25:382–385
28. Beaton DE (2000) Understanding the relevance of measured change through studies of responsiveness. Spine 25(24):3192–3199
29. Guyatt G, Walter S, Norman G (1987) Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 40(2):171–178
30. Ware-JE J, Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 30(6):473–483
31. Jenkinson C, Stewart-Brown S, Petersen S, Paice C (1999) Evaluation of the SF-36 version II in the United Kingdom. J Epidemiol Community Health 53:46–50
32. The Euroqol Group (1990) Euroqol—a new facility for the measurement of health-related quality of life. Health Policy 16:199–208
33. Lim JT, Luscombe KL, Jones PW, White SH (2006) The effect of preoperative symptom severity on functional outcome of total knee replacement–patients with the lowest preoperative scores achieve the lowest marks. Knee 13(3):216–219
34. Hajat S, Fitzpatrick R, Morris R, Reeves B, Rigge M, Williams O, Murray D, Gregg P (2002) Does waiting for total hip replacement matter? Prospective cohort study. J Health Serv Res and Policy 7(1):19–25
35. Vickers AJ, Altman DG (2001) Analysing controlled trials with baseline and follow-up measurements. BMJ 323:1123–1124
36. Porter P, Venkateswaran B, Stephenson H, Wray CC (2002) The influence of age on outcome after operation for the carpel tunnel syndrome. J Bone Joint Surg (Br) 84-B(5):688–691
37. Bremner-Smith AT, Ewings P, Weale AE (2004) Knee scores in a 'Normal' elderly population. Knee 11(4):279–282
38. Jenkinson C, Heffernan C, Doll H, Fitzpatrick R (2006) The Parkinson's Disease questionnaire: evidence for a method of imputing missing data. Age Ageing 35:497–502
39. Bland JM, Altman DG (1996) Transforming data. BMJ 312(7033):770
40. Jaeschke R, Singer J, Guyatt GH (1989) Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 10(4):407–415
41. Wyrwich KW, Tierney WM, Wolinsky FD (1999) Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 52(9):861–873
42. Fayers PM, Machin D (2000) Quality of life—assessment, analysis and interpretation. Wiley, Chichester
43. Norman GR, Sloan JA, Wyrwich KW (2003) Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 41:582–592