EXTENDED REPORT

Interrater reproducibility of clinical tests for rotator cuff lesions

A J K Ostor, C A Richards, A T Prevost, B L Hazleman, C A Speed

Ann Rheum Dis 2004;**63**:1288–1292. doi: 10.1136/ard.2003.014712

Appendix 1 containing additional tables is available at http://www.annrheumdis.com/supplemental

See end of article for authors' affiliations

Correspondence to: Dr A J K Ostor, Rheumatology Research Unit, Box 194, Hills Road, Cambridge CB2 2QQ, UK; andrew.ostor@addenbrookes.nhs.uk

Accepted 1 December 2003

Background: Rotator cuff lesions are common in the community but reproducibility of tests for shoulder assessment has not been adequately appraised and there is no uniform approach to their use.
Objective: To study interrater reproducibility of standard tests for shoulder evaluation among a rheumatology specialist, rheumatology trainee, and research nurse.
Methods: 136 patients were reviewed over 12 months at a major teaching hospital. The three assessors examined each patient in random order and were unaware of each other's evaluation. Each shoulder was examined in a standard manner by recognised tests for specific lesions and a diagnostic algorithm was used. Between-observer agreement was determined by calculating Cohen's κ coefficients (measuring agreement beyond that expected by chance).
Results: Fair to substantial agreement was obtained for the observations of tenderness, painful arc, and external rotation. Tests for supraspinatus and subscapularis also showed at least fair agreement between observers. 40/55 (73%) κ coefficient assessments were rated at >0.2, indicating at least fair concordance between observers; 21/55 (38%) were rated at >0.4, indicating at least moderate concordance between observers.
Conclusion: The reproducibility of certain tests, employed by observers of varying experience, in the assessment of the rotator cuff and general shoulder disease was determined. This has implications for delegation of shoulder assessment to nurse specialists, the development of a simplified evaluation schedule for general practitioners, and uniformity in epidemiological research studies.

Shoulder disorders are common and cause significant morbidity and disability.[1–3] The annual incidence (episodes for which general practitioners are consulted) of shoulder pain has been estimated to be >1% in adults over the age of 45 years,[4] and the prevalence in population studies lies between 7 and 21%.[5 6] Lesions of the rotator cuff account for most episodes of shoulder pain (up to 70%).[6] Lack of consensus about appropriate diagnostic criteria for shoulder disorders, however, plagues research in this area. In a review by Buchbinder *et al*, none of the classification systems for soft tissue disorders of the neck and upper limb appeared acceptable.[7] This lack of consensus accounts, to a large degree, for the paucity of evidence based management approaches in the treatment of rotator cuff lesions.[8]

The need to develop consensus criteria for upper limb disorders has been recognised by the UK Health and Safety Executive[9] and has been a major step forward in the development of structured protocols for the evaluation of specific soft tissue disorders. The criteria benefit from consensus backing but have not yet been fully validated. The schedule includes some limited clinical tests of the shoulder. Palmer and colleagues have contributed to the schedule by evaluating the criteria further and by adding items to form the Southampton Examination Schedule for the diagnosis of musculoskeletal disorders of the upper limb.[10] They found that the schedule was sufficiently reproducible for epidemiological research in the general population.[11]

Despite these steps there is no universally recognised method for evaluating upper limb disorders, with over 20 clinical tests recommended for assessing the rotator cuff alone.[12] None of these tests have been adequately validated in a general population and their sensitivity and specificity are unknown. As a consequence, patients selected for rotator cuff trials are likely to be a heterogeneous group because inclusion is based on clinical criteria. Further difficulties encountered in shoulder studies are problems with case definition,[13] variable experience of examiners,[14] variations in method of assessment, interpretation of physical signs, and diversity of diagnostic categories used.[15–19] This has contributed to conflicting results in interobserver agreement studies, varying from excellent to poor correlation.[20–23]

To be applicable for widespread use, a test must be accurate and reproducible by clinicians with a spectrum of experience. Our study was undertaken to determine the reproducibility of clinical tests for rotator cuff lesions assessed by clinicians with varying levels of experience. The results would then inform which tests had the closest reproducibility and interrater agreement and, therefore, were the most useful.

METHODS

Selection of clinical tests

The clinical tests outlined in the Southampton schedule formed the basis of the examination series used in the study.[10] Additional tests commonly used for assessment of the rotator cuff were included (table 1). The assessors consisted of a consultant rheumatologist with a particular interest in shoulder disorders, a specialist registrar in rheumatology without specific training in shoulder complaints, and a research nurse with no formal musculoskeletal training. The assessment of the consultant was deemed the benchmark against which other assessments were compared.

Sample population

The study group consisted of consecutive patients with shoulder pain referred to the rheumatology unit at a teaching hospital, as well as patients identified in the rheumatology outpatients department whose chief complaint was shoulder pain. The patients were then assessed in a second weekly clinic held at the hospital. Inclusion criteria for participation in the study included ability to give informed consent; age 20–85 years; and shoulder pain regardless of possible

Table 1 Tests used in shoulder examination

Test	Description	Algorithm for diagnosis
Empty can test for supraspinatus	The shoulder is abducted to 90° then internally rotated and brought into 30° forward flexion by the examiner, with thumb pointing downwards. The patient abducts the arm against the examiner's resistance	Pain indicates tendinitis/tendinosis Weakness out of proportion to pain indicates tear
Resisted external rotation for infraspinatus	External rotation resisted with the patient's arm at the side, externally rotated 20° and the elbow flexed to 90°	Pain indicates tendinitis/tendinosis Weakness out of proportion to pain indicates tear
Lift off test for subscapularis	The dorsal aspect of the hand is placed on the ipsilateral buttock and the hand is then lifted off the buttock 1–2″. The hand is then lifted further against the resistance applied by the examiner	Pain indicates tendinitis/tendinosis Weakness out of proportion to pain indicates tear
Yergason's test	With the arm by the side and the forearm flexed to 90° the forearm is supinated against resistance	Positive test: pain in region of bicipital groove indicates tendinitis
Speeds test	With the elbow fully extended and the arm in 30° of flexion further flexion is resisted by the examiner	Positive test: pain in region of bicipital groove indicates tendinitis
Hawkins-Kennedy impingement test	With the patient standing the arm is abducted to 90° and forward flexed to 45°. The arm is then forcibly internally rotated	Positive test elicits pain
Acromioclavicular joint assessment	With patient seated the examiner passively adducts the arm at 90° abduction across the chest	Pain at AC joint indicates pathology (sprain or OA)
	Alternative test: with the patient standing the examiner passively adducts the extended arm in front of the body	Pain at AC joint indicates pathology (sprain or OA)
Drop arm test for rotator cuff rupture	The examiner passively abducts the arm to 90° with subsequent active adduction	Positive test: subject is unable to maintain abduction. Indicates rotator cuff rupture

underlying aetiology. All patients who were referred were offered appointments; however, a small number cancelled before review as their symptoms had improved. Patient characteristics of this group are unknown.

Training

Before starting the trial, two consultant-led training sessions for the other assessors occurred over consecutive weeks. Ten patients were examined by a method similar to that described by Palmer *et al*.[10] The sessions aimed at ensuring that the observing assessors were familiar with the clinical tests involved in the examination series and how to perform them. A handbook, with full details of all clinical tests as they have been reported, was given to each observer. "Training patients" comprised predominantly those with rotator cuff tendinopathy as formally diagnosed by the consultant. Each observer examined the patient under the supervision of the trainer.

Shoulder assessment

The assessment involved three phases, starting with historical information obtained by the first observer. This was followed by inspection of the shoulder for signs of swelling, deformity, tenderness, winging, degree of external rotation, and a "painful arc". Finally, assessment of each part of the rotator cuff was undertaken by resistance testing and was defined as either normal, weakness>pain (implying tear), or pain>weakness (implying tendinosis/tendinitis).[19] Other tests involved assessment of the long head of biceps (Speed and Yerguson tests[12]), signs of impingement using the Hawkins-Kennedy test, and assessment of the acromioclavicular joint (table 1). A provisional diagnosis was made using a predefined algorithm, depending upon the history and examination findings. More than one final diagnosis was possible using this algorithm (table 1).

Table 2 Baseline characteristics of study subjects

Characteristics	
Total subjects (n)	136
Men, No (%)	66 (49)
Women, No (%)	70 (51)
Age (years), mean (range)	
Total	55
Men	53 (20–70)
Women	58 (20–85)
Affected shoulders (n)	
Total	159
Right **only**	64
Left **only**	49
Both left and right	23
Unaffected shoulders	113
Dominant arm (n)	
Right	123
Left	12
Both	1
Duration of symptoms (months)	
Mean (range)	22.9 (1–144)
Median	12

Statistical methods

Between-observer reproducibility of physical signs of shoulder disease was evaluated using Cohen's κ coefficient for categorical variables.[24] Cohen's κ is the preferred statistic for the evaluation of concordance between two clinicians for nominal categories measuring agreement beyond that expected to occur by chance. Landis and Koch proposed guidelines for the interpretation of the strength of concordance reflected by the κ coefficient: <0.00 "poor"; 0.00–0.20 "slight"; 0.21–0.40 "fair"; 0.41–0.60 "moderate"; 0.61–0.80 "substantial"; 0.81–1.00 "almost perfect".[25] A minimum sample size of 125 was chosen on the basis that an observed κ coefficient for a test would then have a 95% confidence interval that covered at most one of these concordance categories either side of the coefficient, on the assumption that the tests would have a test prevalence lying between 20 and 80%.

It is accepted that the sample size will not allow reproducibility of tests for the rare lesions—for example, subscapularis tendinopathies, to be evaluated with precision. Systematic bias between observers was examined with McNemar's test. We did not evaluate intraobserver agreement.

RESULTS

In total, 136 patients were enrolled in the study over a 12 month period. Table 2 shows their demographic information. The number of affected shoulders was 159 with 23 subjects contributing two sets of observations for assessment; 113 shoulders were normal. The analysis assumes independent observations; however, the patients who contributed

Table 3 Agreement between the three clinicians about the presence of tenderness, painful arc, and external rotation in affected shoulders

	Prevalence (%)†			κ Value		
Observations	Consultant	Spec reg	Nurse	Cons/spec reg	Cons/nurse	Spec reg/nurse
Tenderness	68 (n = 159)	55 (n = 159)	58 (n = 159)	0.32**‡	0.26**	0.48**
Painful arc 1 (start of pain)	82 (n = 159)	88 (n = 159)	85 (n = 159)	0.48**‡	0.50**	0.49**
Painful arc 2 (end of pain)	83 (n = 159)	89 (n = 159)	86 (n = 159)	0.64**‡	0.61**	0.62**
External rotation (<45°)	25 (n = 158)	28 (n = 159)	45 (n = 158)	0.68**	0.42**‡	0.46**‡

*p<0.05, **p<0.01; †prevalence is the percentage of affected shoulders which were found to test positive; ‡McNemar p<0.05 denoting significant difference in prevalence between the two raters.

both shoulders represent a relatively small percentage of subjects and are unlikely to have an effect on the results.

κ Coefficients were calculated for all phases of the assessment. All tests were scored on a dichotomous scale. Fair to substantial agreement was obtained for the observations of tenderness, painful arc, and for external rotation (table 3). Agreement between raters was reduced owing to the lower level of prevalence of tenderness rated by the specialist registrar and nurse and the higher level of prevalence of external rotation rated by the nurse.

Tests for supraspinatus and subscapularis also showed at least fair agreement between observers (table 4). Tests for the acromioclavicular joint showed the poorest reproducibility (table 4). Adequate rater agreement in these tests for the overall prevalence among the three raters and agreement in an individual patient was not achievable.

According to the algorithm, diagnosis of shoulder pain showed at least fair concordance for rotator cuff disorders, adhesive capsulitis, and for referred pain, and this was largely associated with a tendency for overestimation of the prevalence of diagnoses by the nurse and specialist registrar relative to consultant. The diagnosis of impingement and acromioclavicular joint disease between consultant and nurse showed only slight agreement (table 5) (fig 1). When the patients were stratified by age, duration of symptoms, sex, and night pain, no improvement in reproducibility for a diagnosis of impingement was found (table 6). In total 40/55 (73%) of the κ coefficient assessments were rated at >0.2, indicating at least fair concordance between observers; 21/55 (38%) were rated at >0.40, indicating at least moderate concordance between observers. There was almost total agreement between observers in assessment of the unaffected shoulder. More detailed analysis of the agreement between observers with full 2×2 tables is available (see Appendix 1 at http://www.annrheumdis.com/supplemental).

DISCUSSION

The results of our study show slight to substantial agreement among observers of varying experience in assessment of the rotator cuff in patients attending a hospital clinic. This is comparable with most previous studies looking at inter-observer agreement of shoulder disorders, although the correlation between research nurse, registrar, and consultant, to our knowledge, has not previously been studied. In addition, earlier studies were not designed to look specifically at the rotator cuff, which represents a narrower field. The results suggest that, with training, delegation of assessment of the shoulder to health staff with no particular expertise may be possible as the optimal agreement for diagnosis was between the specialist registrar and research nurse. This could be useful for epidemiological studies as well as for triage purposes in primary care to maximise use of resources.[11]

In a study by de Winter *et al*, reproducibility was moderate (κ = 0.45) between two physiotherapists assessing shoulder disorders.[26] The poorest diagnostic agreement was found for subjects with a high degree of pain, bilateral involvement, and chronic complaints. Their patient population was more selective than in our study and therefore our results might have improved if we had not included patients with shoulder pain regardless of possible underlying cause. One explanation for the diagnostic disagreement in the study of de Winter *et al* was the insufficient mutual exclusivity of diagnostic categories currently used for shoulder disorders—an example being pain associated with adhesive capsulitis rendering any further assessment of the shoulder difficult. In the study of Liesdek *et al*,[20] in contradistinction to that of de Winter *et al*,[26] agreement between general practitioners and physiotherapists in diagnosis of soft tissue disorders of the shoulder was greater in patients with a long duration of symptoms. Their overall κ value for the classification of shoulder disorders was fair at 0.31 and might have been an overestimation, as the physiotherapists were not "blinded" to the diagnosis of the general practitioners.

Bamji *et al* showed poor agreement between consultant rheumatologists in shoulder assessment, with agreement in only 46% when examined independently and 78% when assessed together.[21] One study has, however, shown very high agreement between physiotherapists,[22] which is in contrast

Table 4 Agreement between three clinicians about resistance testing in the affected shoulder

	Prevalence (%)†			κ Value		
Tests	Consultant	Spec reg	Nurse	Cons/spec reg	Cons/nurse	Spec reg/nurse
Supraspinatus (empty can test)	68 (n = 157)	79 (n = 158)	75 (n = 158)	0.49**‡	0.44**	0.46**
Infraspinatus (resisted external rotation)	44 (n = 156)	47 (n = 158)	52 (n = 159)	0.45**	0.18*	0.38**
Subscapularis (lift off test)	44 (n = 156)	54 (n = 158)	55 (n = 159)	0.30**	0.32**‡	0.28**
Long head of biceps (Yergason's test)	0 (n = 158)	9 (n = 157)	16 (n = 157)	N/A‡§	N/A‡§	0.276**‡
Long head of biceps (Speeds test)	13 (n = 158)	25 (n = 159)	30 (n = 159)	0.17*‡	0.26**‡	0.32**
Impingement (Hawkins-Kennedy)	49 (n = 152)	73 (n = 158)	68 (n = 156)	0.29**‡	0.18*‡	0.43**
Acromioclavicular joint (elevation)	14 (n = 134)	30 (n = 148)	50 (n = 149)	0.19*‡	0.20**‡	0.30**‡
Acromioclavicular joint (adduction front)	7 (n = 141)	26 (n = 152)	42 (n = 158)	0.09‡	0.08‡	0.29**‡
Acromioclavicular joint (adduction back)	1 (n = 138)	25 (n = 147)	52 (n = 142)	0.06‡	0.03‡	0.38**‡
Drop arm test	0.8 (n = 126)	4 (n = 129)	2 (n = 135)	0.275**	0.659**	0.528**

*p<0.05, **p<0.01; †prevalence is the percentage of affected shoulders which were found to test positive; ‡McNemar p<0.05 denoting significant difference in prevalence between the two raters; §not applicable owing to a prevalence of zero.

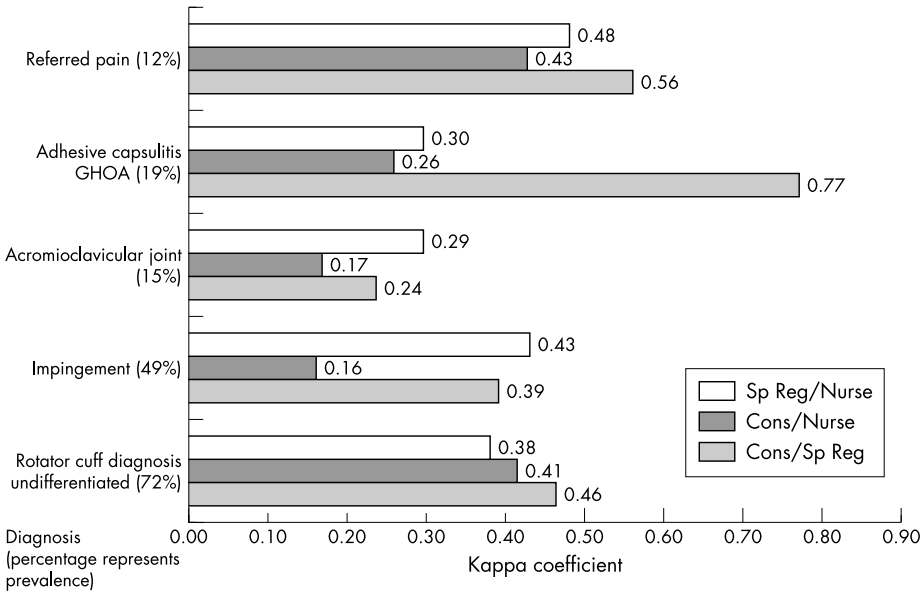

Figure 1 Agreement among three clinicians in diagnosis in affected shoulders assessed by κ coefficient.

with the other interobserver studies. The limitations of this study were the small numbers (19 patients), the paucity of information on the clinical characteristics of the patients, and the setting of the study.

The difficulty in assessing soft tissue disorders of the shoulder is further highlighted by observer variability in shoulder movement. Croft *et al* found that the assessment of external rotation was poorly reproducible owing to random variation in visual estimation and systematic variation in examination technique.[27] In a study by Hoving *et al*, adequate reliability was only obtained for the movements of total shoulder flexion and hand behind back.[28]

Our results show that there was at least moderate agreement for lesions of the supraspinatus muscle and for signs of capsulitis. A limitation of the κ value, when the background prevalence of the parameter in question is particularly low or high, is that κ coefficients have been shown to be smaller than when the prevalence is in the middle of the range (50%) when the sensitivity and specificity of the test are kept the same. Another limitation with extreme prevalence is that the κ coefficient has relatively large sample to sample variability and wide confidence intervals. Taken together, these limitations may partially explain the reduced reproducibility in tests for less common lesions, which would require evaluation in a larger sample. Differences in prevalence between study populations also complicate effective comparison of κ coefficients between studies because the observed value of κ depends on the similarity of prevalence in the studies.[29]

It is axiomatic that the true worthiness of a diagnostic test lies in its ability to alter treatment and/or in prognostication. There is unfortunately scant evidence about the usefulness of clinical tests for shoulder disease in either of these areas. As the reproducibility of many of the tests in our study was poor, this hampers further efforts to use a large set of tests by various practitioners to aid in shoulder pain management. This issue is further complicated when assessment is undertaken in general practice, where the presentation or severity of disease may be quite different from that seen in a teaching hospital. The minimum set of tests we found to have at least moderate reproducibility comprised those for painful arc, external rotation, and empty can. This is a reasonable start as these tests imply lesions of the capsule (such as frozen shoulder, which is notoriously recalcitrant to treatment) and supraspinatus, the most commonly affected muscle in rotator cuff lesions and hence the main culprit causing shoulder pain.

Although we focused principally on rotator cuff lesions, rarer diagnoses encountered, such as instability, reflex sympathetic dystrophy, myofascial pain, and fibromyalgia, were not included in the diagnostic algorithm. The research nurse, who had no previous knowledge of shoulder disorders, could not make these diagnoses. This highlights one the limitations of this study and the difficulties in standardising the examination. Missing rare diagnoses by less experienced staff may not be important, however, as patients with an unclear initial assessment could be referred for specialist opinion. We felt a diagnostic algorithm was necessary in order to translate the clinical signs into an entity which could be used for treatment.

The usual management of rotator cuff lesions, although not adequately validated[8] but employed by clinicians,

Table 5 Agreement between the three clinicians about diagnosis of affected arm

	Prevalence (%)†			κ Value		
Diagnosis	Consultant	Spec reg	Nurse	Cons/spec reg	Cons/nurse	Spec reg/nurse
Rotator cuff diagnosis (undifferentiated)	72 (n=153)	82 (n=158)	82 (n=159)	0.46**‡	0.41**‡	0.38**
Impingement	49 (n=156)	69 (n=159)	66 (n=159)	0.39**‡	0.16*‡	0.43**
Acromioclavicular joint	15 (n=156)	27 (n=159)	46 (n=159)	0.24**‡	0.17**‡	0.29**‡
Adhesive capsulitis/GHOA	19 (n=156)	23 (n=159)	33 (n=159)	0.77**	0.26**‡	0.30**‡
Referred pain	12 (n=156)	7 (n=159)	14 (n=159)	0.56**‡	0.43**	0.30**‡

*p<0.05, **p<0.01; †prevalence is the percentage of affected shoulders which were found to test positive; ‡McNemar p<0.05 denoting significant difference in prevalence between the two raters.

Table 6 Agreement between clinicians about assessment of impingement related to age, sex, duration of symptoms, and presence of night pain

Diagnosis of impingement	Consultant/ Spec reg	Consultant/ Nurse	Spec reg/ Nurse	Total
Age <55 years	0.40	0.26	0.45	70
Age ⩾55 years	0.19	0.1	0.34	81
Duration ⩽12 weeks	0.13	0.15	0.37**	83
Duration >12 weeks	0.53**	0.23	0.49**	68
In men	0.28**	0.23**	0.57**	75
In women	0.32**	0.13	0.30**	76
With night pain	0.30**	0.16**	0.38**	138
With no night pain	0.24	0.40	0.84**	13

*p<0.05, **p<0.01.

involves a variable combination of analgesics, physiotherapy, and injection of corticosteroid and local anaesthetic into the subacromial space or glenohumoral joint. This may be initiated in primary care, with referral of the more severe or unclear cases to specialist clinics. Patients who undergo this treatment paradigm who do not improve could also be referred for specialist assessment.

Despite an attempt at standardisation made in the Southampton Examination Schedule, diagnosis of rotator cuff pathology was the most difficult and the least defined. Training of the observers lasted for 6 weeks until consistency was optimised, compared with two teaching sessions in our study, making it less applicable to daily practice. Our results will have to be repeated, however, to see if the reproducibility holds across all practitioners who deal with shoulder disorders and in primary care where subjects may have less defined illness. Our results might have improved with a longer duration and more intensive training session.

Validation of specific clinical tests requires a standard of reference which in many cases may not be available or easily accessible. In our study the consultant's opinion was taken as the benchmark, as has been used in other studies[10]; however, this has limitations as true validity is not assessed. The consultant's opinion was deemed appropriate and adequate, however, for a reproducibility study of this nature.

Primary care is the most appropriate location for appraisal of shoulder disorders. Direct orthopaedic referral in many institutions is becoming increasingly difficult, with the implication of increased referral to the rheumatology department. If most shoulder disorders could be adequately managed in primary care this would significantly reduce the tertiary referral workload.

In summary, this study has identified the reproducibility of certain tests, employed by observers of varying experience, in the assessment of the rotator cuff and general shoulder disease. Further study of these tests is required across all disciplines involved with shoulder disease. The results have important implications for future epidemiological and treatment studies.

ACKNOWLEDGEMENTS

Consulting services at the Centre for Applied Medical Statistics, University of Cambridge provided statistical support.

Authors' affiliations

A J K Ostor, C A Richards, A T Prevost, B L Hazleman, C A Speed, Rheumatology Research Unit, Addenbrooke's Hospital, Cambridge, UK

REFERENCES

1 **Croft PR**, Pope DP, Silman A. The clinical course of shoulder pain:prospective cohort study in primary care. *BMJ* 1996;**313**:601–2.
2 **van der Windt DA**, Koes BW, Boeke AJ, Deville W, De Jong BA, Bounter LM. Shoulder disorders in primary care: prognostic indicators of outcome. *Br J Gen Pract* 1996;**46**:519–23.
3 **Macfarlane GJ**, Hunt IM, Silman AJ. Predictors of chronic shoulder pain: a population based prospective study. *J Rheumatol* 1998;**25**:1612–15.
4 **Royal College of General Practitioners**. Office of Population Census and Surveys, Department of Health and Social Security. *Morbidity statistics from general practice. Third national study:socio-economic analyses*. London: HMSO, 1986 (Series MB5 No 2).
5 **Bjelle A**. Epidemiology of shoulder problems. *Baillieres Clin Rheumatol* 1989;**3**:437–51.
6 **Chard MD**, Hazleman R, Hazleman BL, King RH, Reiss BB. Shoulder disorders in the elderly: a community survey. *Arthritis Rheum* 1991;**34**:766–9.
7 **Buchbinder R**, Goel V, Bombardier C, Hogg-Johnson S. Classification systems of soft tissue disorders of the neck and upper limb: do they satisfy methological guidelines? *J Clin Epidemiol* 1996;**49**:141–9.
8 **Green S**, Buchbinder R, Glazier R, Forbes A. Interventions for shoulder pain (Cochrane review). In: *The Cochrane Library*. Oxford: Update Software, 2002;(3).
9 **Harrington JM**, Carter JT, Birrell L, Gompertz D. Surveillance case definitions for work related upper limb pain syndromes. *Occup Environ Med* 1998;**55**:264–71.
10 **Palmer K**, Walker-Bone K, Linaker C, Reading I, Kellingray S, Coggon D, *et al*. The Southampton examination schedule for the diagnosis of musculoskeletal disorders of the upper limb. *Ann Rheum Dis* 2000;**59**:5–11.
11 **Walker-Bone K**, Byng P, Linaker C, Reading I, Coggon D, Palmer K, *et al*. Reliability of the Southampton examination schedule for the diagnosis of upper limb disorders in the general population. *Ann Rheum Dis* 2002;**61**:1103–6.
12 **Magee DJ**. Shoulder. In: *Orthopaedic physical assessment*. 2nd ed. Philadelphia: Saunders 90–142.
13 **Pope DP**, Croft PR, Pritchard CM, Silman AJ. Prevalence of shoulder pain in the community: the influence of case definition. *Ann Rheum Dis* 1997;**56**:308–12.
14 **Balich SM**, Sheley RC, Brown TR, Sauser DD, Quinn SF. MR imaging of the rotator cuff tendon: interobserver agreement and analysis of interpretative errors. *Radiology* 1997;**204**:191–4.
15 **Neer CS**. Impingement lesions. *Clin Orthop* 1983;**173**:70–7.
16 **Uhthoff HK**, Sarkar K. An algorithm for shoulder pain caused by soft-tissue disorders. *Clin Orthop* 1990;**254**:121–7.
17 **Hedtmann A**, Fett H. The so-called periarthropathic humeroscapularis. Classification and analysis of 1266 cases. *Z Orthop* 1989;**127**:643–9.
18 **Bakker JF**, de Jongh L, Jonquière M, Mens J, Oosterhun WW, Poppelaars A, *et al*. Standaard Schouderklachten. *Huisarts en Wetenschap* 1990;**33**:196–202.
19 **Cyriax J**. *Textbook of orthopaedic medicine*, 7th ed. London: Baillière Tindall, 1981:190–239.
20 **Liesdek C**, van der Windt DAWM, Koes BW, Bouter LM. Soft-tissue disorders of the shoulder: a study of inter-observer agreement between general practitioners and physiotherapists and an overview of physiotherapeutic treatment. *Physiotherapy* 1997;**83**:12–17.
21 **Bamji AN**, Erhardt CC, Price TR, Williams PL. The painful shoulder: can consultants agree? *Br J Rheumatol* 1996;**35**:1172–4.
22 **Pellecchia GL**, Paolino J, Connell J. Intertester reliability of the Cyriax evaluation in assessing patients with shoulder pain. *J Orthop Sports Phys Ther* 1996;**23**:34–8.
23 **Nørregaard J**, Krogsgaard MR, Lorenzen T, Jensen EM. Diagnosing patients with longstanding shoulder joint pain. *Ann Rheum Dis* 2002;**61**:646–9.
24 **Cohen J**. A coefficient of agreement for nominal scales. *Educ Psychol Measure* 1960;**20**:37–46.
25 **Landis JR**, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
26 **de Winter AF**, Jans MP, Scholten RJPM, Devillé W, van Schaardenburg D, Bounter LM. Diagnostic classification of shoulder disorders: interobserver agreement and determinants of disagreement. *Ann Rheum Dis* 1999;**58**:272–7.
27 **Croft P**, Pope D, Boswell R, Rigby A, Silman A. Observer variability in measuring elevation and external rotation of the shoulder. *Br J Rheumatol* 1994;**33**:942–6.
28 **Hoving JL**, Buchbinder R, Green S, Forbes A, Bellamy N, Brand C, *et al*. How reliably do rheumatologists measure shoulder movement? *Ann Rheum Dis* 2002;**61**:612–16.
29 **Thompson WD**, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol* 1988;**41**:949–58.