CONCISE REPORT

Reliability of the Southampton examination schedule for the diagnosis of upper limb disorders in the general population

K Walker-Bone, P Byng, C Linaker, I Reading, D Coggon, K T Palmer, C Cooper

Ann Rheum Dis 2002;61:1103–1106

Background: Epidemiological research in the field of soft tissue neck and upper limb disorders has been hampered by the lack of an agreed system of diagnostic classification. In 1997, a United Kingdom workshop agreed consensus definitions for nine of these conditions. From these criteria, an examination schedule was developed and validated in a hospital setting.

Objective: To investigate the reliability of this schedule in the general population.

Methods: Ninety seven adults of working age reporting recent neck or upper limb symptoms were invited to attend for clinical examination consisting of inspection and palpation of the upper limbs, measurement of active and passive ranges of motion, and clinical provocation tests. A doctor and a trained research nurse examined each patient separately, in random order and blinded to each other's findings.

Results: Between observer repeatability of the schedule was generally good, with a median κ coefficient of 0.66 (range 0.21 to 0.93) for each of the specific diagnoses considered.

Conclusion: As expected, the repeatability of tests is poorer in the general population than in the hospital clinic, but the Southampton examination schedule is sufficiently reproducible for epidemiological research in the general population.

Musculoskeletal disorders of the upper limb and neck are common in people of working age and costly in terms of time lost from work.¹ Research into their causes, management, and prevention is thus important. However, investigation in this field has been hampered by the lack of an agreed system of case classification.² As a consequence, differences in taxonomy and the use of diagnostic labels have persisted, despite several attempts at standardisation.^{3–9} Moreover, even the standardised systems that have been developed have been criticised on methodological grounds, in particular for failing to show satisfactory repeatability and validity.¹⁰

To consider these concerns, a multidisciplinary British group was convened in 1997 (the Birmingham workshop) with the objective of deriving consensus criteria for the more common upper limb disorders.¹¹ Based on these criteria, we have developed a standardised examination protocol and evaluated its performance in 88 patients attending hospital rheumatology and orthopaedic outpatient clinics with soft tissue disorders of the upper limb. The new examination system proved to be repeatable between observers (trained research nurses), and had acceptable diagnostic accuracy when compared with the clinical diagnosis of a physician as the reference standard.¹² However, our hospital based study population would have overrepresented the more severe and

clear cut cases of soft tissue musculoskeletal complaints. The objective of this study therefore, was to evaluate the utility of the new examination protocol in the general population, where abnormal findings would be expected to be more subtle and less frequent.

METHODS

The clinical examination was performed in a sample of men and women aged 25–64 years who had participated in a population based cross sectional survey of the prevalence of neck and upper limb complaints.¹³ All 6660 men and women registered with a general practice in Southampton were sent a postal questionnaire about recent pain in the neck or upper limb. Everyone who indicated that they had experienced neck or upper limb pain lasting a day or longer, or numbness or tingling lasting three minutes or longer, in the preceding week (2162 of 3991 respondents) was invited to attend for an interview and physical examination. Among those who attended (1334 patients (62%)), a group of 97 (7%) were asked if they were willing to be examined on a second occasion. The sample of 97 was selected on the basis of attendance on a day of the week when two observers were present (but all working days were used equally). All who were invited agreed to participate.

Each patient was examined independently by a nurse and a rheumatologist, at an interval of a few minutes. Examinations were carried out in random order with each observer unaware of the other's findings. The same rheumatologist examined every patient and the study was designed so that each of two trained research nurses performed about half of the examinations. Each examination took about 15 minutes to perform.

The between observer repeatability for eliciting items in the physical examination was assessed by calculating a Cohen's κ coefficient for categorical variables,¹⁴ and mean differences with limits of agreement for continuous variables.¹⁵ The repeatability of the diagnoses derived from the physical findings of each examiner was also assessed.

RESULTS

Ninety seven patients were included in the study (49 examined by nurse 1 and 48 by nurse 2). The mean age of participants was 48 years (range 29–65 years) and 59 (61%) were women. The prevalence of abnormal physical signs was low in this population sample. The most frequent abnormality was the presence of Heberden's nodes, noted in 105 hands in total (by either or both observers), with the second being Dupuytren's contracture, found in only 35 hands. Table 1 summarises the between observer repeatability with which the two nurses and the rheumatologist elicited individual observations (symptoms and physical signs). κ Coefficients varied widely between different elements of the schedule (–0.03 to 0.94). Values were greater than 0.50 for shoulder pain (κ =0.94); anterior shoulder pain (κ =0.66); acromioclavicular joint pain (κ =0.56); shoulder pain on resisted elbow

Table 1 Between observer repeatability of physical signs examined in the schedule

Symptoms or physical signs	No. of pairs examined	Nurse/rheumatologist				κ (SE)
		-/-	-/+	+/-	+/+	
Shoulder:						
Shoulder pain:						
Deltoid pain	194	175	7	6	6	0.44 [0.07]
Anterior pain	194	188	2	1	3	0.66 [0.07]
Acromioclavicular	194	189	1	2	2	0.56 [0.07]
Subacromial bursa	194	189	1	4	0	-0.01 [0.06]
Any shoulder pain	194	163	1	2	28	0.94 [0.07]
Shoulder pain on resisted:						
Elbow flexion	194	190	0	2	2	0.66 [0.07]
Forearm supination	194	189	3	1	1	0.32 [0.07]
External rotation	194	184	4	3	3	0.44 [0.07]
Internal rotation	194	186	5	3	0	-0.02 [0.07]
Abduction	194	170	9	10	5	0.29 [0.07]
Painful arc	194	179	7	3	5	0.47 [0.07]
Positive AC joint stress test	194	170	6	13	5	0.30 [0.07]
Elbow:						
Elbow tenderness:						
Lateral	194	178	4	6	6	0.52 [0.07]
Medial	194	176	4	5	9	0.64 [0.07]
Lateral elbow pain, resisted extension	194	183	4	3	4	0.52 [0.07]
Medial elbow pain, resisted flexion	194	184	2	4	4	0.56 [0.07]
Wrist/hand:						
Radial wrist tenderness	194	184	4	5	1	0.16 [0.07]
Pain on resisted movement:						
Radial wrist	194	183	6	5	0	-0.03 [0.07]
Medial wrist	194	189	3	2	0	-0.01 [0.07]
Finger extension	194	188	3	1	2	0.49 [0.07]
Finger flexion	194	186	1	5	2	0.49 [0.06]
Thenar muscle wasting	194	192	0	2	0	0.00 [-]
Dupuytren's contracture	194	159	15	0	20	0.69 [0.07]
Heberden's nodes	194	89	24	14	67	0.60 [0.07]
Abnormal light touch:						
Thumb	194	181	5	3	5	0.53 [0.07]
Index finger	194	176	9	6	3	0.25 [0.07]
Little finger	194	178	7	6	3	0.28 [0.07]
Positive Phalen's test	194	165	1	12	16	0.68 [0.07]
Positive Tinel's test	194	182	2	7	3	0.38 [0.07]
Weakness of thumb:						
Abduction	194	190	0	3	1	0.40 [0.06]
Opposition	194	191	0	3	0	0.00 [-]
Pain resisted thumb extension	194	181	1	9	3	0.36 [0.06]
Positive Finkelstein's test	194	177	7	6	4	0.35 [0.07]

Observer 1=nurse; observer 2=rheumatologist.

-/- negative to both observers; -/+ negative to observer 1, positive to observer 2; +/- positive to observer 1, negative to observer 2; +/+ positive to both observers.

flexion ($\kappa=0.66$); tenderness over the lateral ($\kappa=0.52$) and medial elbow ($\kappa=0.64$); lateral elbow pain on resisted wrist extension ($\kappa=0.52$); medial elbow pain on resisted wrist flexion ($\kappa=0.56$); Dupuytren's contracture ($\kappa=0.69$); Heberden's nodes ($\kappa=0.60$); a positive Phalen's test ($\kappa=0.68$); and abnormal light touch in the thumb ($\kappa=0.53$).

Measurements of the range of shoulder and neck movement showed a high level of concordance between observers, with the most repeatable being internal rotation of the shoulder (mean difference -0.1° , limits of agreement -0.3° to $+0.1^\circ$) and the least being abduction of the shoulder (mean difference -9° ; limits of agreement -11° to -7°).

Individual physical findings were combined to yield specific diagnoses using the consensus criteria of the Birmingham workshop.¹¹ Table 2 shows the repeatability of classification into these diagnostic categories using the physical findings from the two observers. κ Coefficients again varied widely (0.21 to 0.93). Poorest agreement was found for the diagnosis of hand/wrist tenosynovitis ($\kappa=0.21$), rotator cuff tendinitis ($\kappa=0.35$), and adhesive capsulitis ($\kappa=0.39$). There was greater concordance for the diagnoses of bicipital tendinitis ($\kappa=0.49$),

medial epicondylitis ($\kappa=0.66$), De Quervain's tenosynovitis ($\kappa=0.66$), lateral epicondylitis ($\kappa=0.75$), and carpal tunnel syndrome ($\kappa=0.93$).

The Birmingham consensus criteria for the diagnosis of rotator cuff tendinitis and adhesive capsulitis specify that pain must be in the "deltoid" region in the presence of characteristic physical signs.¹¹ Table 1 shows that the reproducibility of reporting pain in the deltoid region was poorer than that for the whole shoulder. We therefore reanalysed our data using the more general criterion of "shoulder pain" in the presence of the characteristic physical signs of rotator cuff tendinitis and adhesive capsulitis (table 2). The κ coefficient for adhesive capsulitis improved to 0.66 and that for rotator cuff tendinitis improved to 0.46.

Analysis was repeated separately for each nurse versus the rheumatologist (data not presented), but no systematic differences were found. In addition, analyses which adjusted for the nesting of observations within patients were conducted, but these made only slight differences to the estimated κ values (≤ 0.11).

Table 2 Between observer repeatability of clinical diagnoses using the schedule

Diagnosis	No.	Nurse v rheumatologist				κ
		-/-	+/-	-/+	+/+	
Adhesive capsulitis	194	179	5	6	4	0.39
Rotator cuff tendinitis	194	181	3	7	3	0.35
Bicipital tendinitis	194	191	1	1	1	0.49
Lateral epicondylitis	194	189	2	0	3	0.75
Medial epicondylitis	194	190	2	0	2	0.66
De Quervain's tenosynovitis	194	192	0	1	1	0.66
Carpal tunnel syndrome	194	168	2	1	23	0.93
Tenosynovitis	194	186	2	5	1	0.21
Revised definition of "shoulder" pain:*						
Adhesive capsulitis	194	166	9	4	15	0.66
Rotator cuff tendinitis	194	173	8	6	7	0.46

* See text.

DISCUSSION

Sound epidemiological research depends on a reliable system of defining and classifying cases. This is particularly problematic for soft tissue rheumatism, given the absence of a clear cut gold standard, and those classification schemes that have been developed so far³⁻⁹ have been criticised on methodological grounds.¹⁰

Prerequisite criteria for a satisfactory scheme include face and content validity, repeatability, and predictive validity.¹⁰ In this respect the Southampton examination schedule has several advantages. Firstly, it was developed after a workshop of experts from many disciplines, and is supported by clinical consensus.¹¹ (Similar criteria to those of the Birmingham workshop have also been developed in The Netherlands, thereby widening the extent of consensus.¹⁶) Secondly, it has been tested previously in the hospital setting and was found to be repeatable, with an acceptable diagnostic accuracy relative to a specialist clinic's independent opinion.¹² Thirdly, the practical feasibility of the schedule is well established. To this list may now be added the evidence presented in this paper on its repeatability in a community setting.

Generally, the between observer repeatability was found to be poorer in the general population than in the hospital clinic. According to Fleiss' criteria, a κ of 0.2 denotes fair, and a κ of 0.4 to 0.7 good agreement.¹⁴ On this basis, 18 of the variables in table 1 had good reproducibility, but for 15 it was only fair. Those who present in secondary care with specific upper limb disorders are likely to represent a severely affected group and it might be hypothesised that physical signs in the general population would be less clear cut and more difficult to detect reliably. However, the signs were not all of equal importance to diagnosis, and at this level the performance of the schedule was better ($\kappa \geq 0.46$ for seven of eight diagnoses).

The prevalence of abnormality was low in this population sample—a situation in which it is easier for observers to agree by chance alone. The κ statistic is constructed to correct for this effect, and demands greater agreement between observers for a given κ value when the prevalence is low (or high) than when measuring the repeatability of an abnormality with a prevalence near 50%.

The study sample was drawn from patients who attended a first assessment in a community survey of upper limb disorders. Everyone who was asked to undergo a second examination agreed to do so (participation rate 100%), but the sampling frame represented a subset of those to whom a questionnaire had been mailed originally. If only severely affected patients agreed to attend the first assessment (spectrum bias), then this circumstance would tend to favour agreement on observations between observers. But, as judged by their postal response, patients from the community survey

who attended for interview were no more likely to have reported disabling pain than those who were invited but did not attend (data not presented), so biased assessment of between observer repeatability is unlikely to have arisen in this way.

The advantages of the schedule need to be considered alongside certain limitations. Our findings on repeatability followed a period of extensive training (about 10–12 sessions) and periodic checking (6–12 monthly between observer studies) of the research nurses, to promote consistency between observers. It should be remembered that agreement between observers is not a property that is fixed but that it can be improved by careful attention to induction and refresher training. In some areas of weak repeatability, such as diagnosis of tenosynovitis at the wrist, the need for better training was highlighted by our data. The difficulty was not found to be in the technical procedure of performing the tests, but in the recognition of abnormal findings. Both nurses had performed significant numbers of examinations, but had seen few cases of wrist tenosynovitis, so that they found the physical signs of swelling of tendon sheaths and pain on resisted movement difficult to discriminate. In this context, it is noteworthy that nurses in training should be given enough opportunity to practise all the tests in sufficient numbers of abnormal cases. Also, there were systematic differences between the nurses and rheumatologist for some outcomes (for example, there were no cases of Dupuytren's contracture in which the nurse made the diagnosis and the rheumatologist did not, but the converse happened; on the other hand, the nurse was more likely to identify a positive Phalen's test than the rheumatologist). These circumstances again point to the need for careful and sufficient training.

To derive a workable examination protocol, the consensus statements of the Birmingham workshop were converted into a detailed diagnostic algorithm that defined anatomical regions and specified how to perform clinical tests. To do this, selections were made (from among a range of possible choices) aiming to maximise agreement and repeatability, but which were compatible with the workshop's judgment. However, our data from the shoulder illustrate the potential pitfalls of this approach. The consensus definitions proposed that rotator cuff tendinitis and adhesive capsulitis were associated with pain in the "deltoid region". Accordingly, we developed a mannequin defining the deltoid region based on the epaulet position of the deltoid muscle. However, we found that patients in the general population were not reliably able to localise their shoulder pain to such a specific site, even when examinations were carried out on the same day, a few moments apart. It has been our observation that between observer repeatability of these diagnoses was considerably improved by loosening the criteria to shoulder pain. Interestingly, most other classification systems have employed the criterion of "shoulder pain" for adhesive capsulitis³⁻⁵ and rotator cuff tendinitis.^{3-5, 8, 9} Although "deltoid pain" was proposed as the anatomical site of rotator cuff tendinitis in two classification systems,^{6, 7} neither study reported the reliability of this criterion either within or between observers.

On balance, although the repeatability of the diagnostic schedule is poorer in the general population than in the hospital clinic, the Southampton examination schedule seems to be sufficiently reproducible for epidemiological research into soft tissue musculoskeletal disorders of the neck and upper limb in the community. The importance of making a diagnosis lies in its utility in distinguishing groups of patients who require different case management, and diagnoses that carry different prognoses or different associations with modifiable risk factors. The predictive validity of a classification scheme is thus a critical test of its usefulness, and this now needs to be evaluated for the Southampton schedule.

ACKNOWLEDGEMENTS

We are grateful to Vanessa Cox and Ken Cox, who provided computer support and to the staff at the MRC Unit, Southampton who helped in data handling. This study was supported by a grant from the Health and Safety Executive and a project grant PO552, from the Arthritis Research Campaign. Infrastructure support was provided by the Medical Research Council. KW-B was supported by an ARC Clinical Research Fellowship and IR by a fellowship grant from the Colt Foundation.

Authors' affiliations

K Walker-Bone, P Byng, C Linaker, I Reading, D Coggon, K T Palmer, C Cooper, MRC Environmental Epidemiology Unit, University of Southampton, UK

Correspondence to: Professor C Cooper, MRC Environmental Epidemiology Unit, Southampton General Hospital, Tremona Road, Southampton, SO16 6YD, UK; gjs@mrc.soton.ac.uk.

Accepted 7 May 2002

REFERENCES

- Bureau of Labor Statistics.** Occupational injuries and illnesses in the United States by industry—1990. Bulletin 2399. Washington DC: US Department of Labor, 1992:42–6.
- Palmer K, Coggon D, Cooper C, Doherty M.** Work related upper limb disorders: getting down to specifics. *Ann Rheum Dis* 1998;57:445–6.
- Waris P, Kuorinka I, Kurppa K, Luopajarvi T, Virolainen M, Pesonen K, et al.** Epidemiologic screening of occupational neck and upper limb disorders. *Scand J Work Environ Health* 1979;53:25–38.
- Viikari-Juntura E.** Neck and upper limb disorders among slaughterhouse workers. *Scand J Work Environ Health* 1983;9:283–90.
- Ohlsson K, Attewell RG, Johnsson B, Ahlm A, Skerfving S.** An assessment of neck and upper limb upper extremity disorders by questionnaire and physical examination. *Ergonomics* 1994;37:891–7.
- Ranney D, Wells R, Moore A.** Upper limb musculoskeletal disorders in highly repetitive industries: precise anatomical physical findings. *Ergonomics* 1995;38:1408–23.
- Silverstein BA.** *The prevalence of upper extremity cumulative trauma disorders in industry*, [thesis]. The University of Michigan, Occupational Health and Safety, 1985.
- McCormack RR, Inman RD, Wells A, Berntsen A, Imbus HR.** Prevalence of tendinitis and related disorders of the upper extremity in a manufacturing workforce. *J Rheumatol* 1990;17:958–64.
- Kaergaard A, Andersen JH.** Musculoskeletal disorders of the neck and shoulders in female sewing machine operators. *Occup Environ Med* 2000;57:528–34.
- Buchbinder R, Goel V, Bombardier C, Hogg-Johnson S.** Classification systems of soft tissue disorders of the neck and upper limb: do they satisfy methodological guidelines? *J Clin Epidemiol* 1996;49:141–9.
- 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.
- 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 neck and upper limb. *Ann Rheum Dis* 2000;59:5–11.
- Walker-Bone, Palmer K, Reading I, Byng P, Shipp A, Linaker C, et al.** A population survey of the frequency of specific and non-specific disorders of the neck and upper limbs. *Arthritis Rheum* 2000;43:S369.
- Cohen S, Karmack T, Mermelstein R.** A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;20:37–46.
- Bland JM, Altman DG.** Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307–10.
- Sluiter JK, Rest KM, Frings-Dresen M.** Criteria document for evaluating the work-relatedness of upper-extremity musculoskeletal disorders. *Scand J Work Environ Health* 2001;27(suppl 1):3–102.