La Medicina del Lavoro Med Lav 2007; *98*, 2: 111-117

The Southampton Examination Schedule for the diagnosis of musculoskeletal disorders of the upper limb

K.T. Palmer
MRC Epidemiology Resource Centre, University of Southampton, Southampton General Hospital, UK

Key words
Musculoskeletal disease; upper extremity; diagnostic criteria

Summary
Background: *The optimum classification of upper limb disorders (ULDs) remains a cause of debate. Recent efforts to address the issue have focused on translating the consensus criteria of experts into workable protocols for use in field epidemiology.* **Objectives:** *This paper describes the development and assessment of one such protocol, the Southampton Examination Schedule for ULDs.* **Results and Conclusions:** *In the absence of a reliable gold standard, the schedule has so far been evaluated in terms of its repeatability within and between-observers in clinical and community settings, and in terms of its capacity to distinguish groups with different severity of disease, different treatment needs, different risk factors and different prognoses. Findings to date are briefly summarised. The most pressing future goal in this field is for researchers to collect data on the component elements of diagnosis according to common evidence-based standards such as the Southampton Schedule in order to facilitate communication, the effective pooling of data and the empirical assessment of alternative choices of case definition.*

Riassunto
«La "Southampton Examination Schedule" per la diagnosi delle patologie muscolo-scheletriche dell'arto superiore». *La più idonea classificazione dei disturbi dell'arto superiore rimane oggi motivo di dibattito. Recenti lavori, rivolti a tale problematica, hanno tentato di tradurre i criteri di consenso dettati dagli esperti in protocolli utilizzabili in campo epidemiologico. Questo articolo descrive lo sviluppo e la valutazione di uno di questi protocolli: "*The Southampton Examination Schedule" *per i disturbi dell'arto superiore. In assenza di un gold standard affidabile, la "*Schedule" *è stata finora valutata in termini di ripetibilità intra-osservatore e tra osservatori diversi, sia in ambito clinico che nella popolazione generale, e in termini di affidabilità nel distinguere gruppi differenti per quel che riguarda la severità della patologia, la necessità di cure, i fattori di rischio e la prognosi. I risultati finora ottenuti vengono brevemente riassunti. È indispensabile per il futuro che un obiettivo prioritario sia l'utilizzo da parte dei ricercatori di strumenti per la raccolta dei dati basati sull'evidenza, come "*The Southampton Examination Schedule"*, al fine di facilitare la comunicazione, di ottenere una banca dati omogenea e di valutare empiricamente possibili alternative nella definizione di caso.*

Pervenuto il 6.12.2006 - Accettato il 3.1.2007
Corrispondenza: Dr. Keith Palmer, MRC Epidemiology Resource Centre, Southampton General Hospital, Tremona Road, Southampton, SO16 6YD UK - Tel. +44 (0) 23 8077 7624 - Fax +44 (0) 23 8070 4021 - E-mail: ktp@mrc.soton.ac.uk

This paper was presented at the ICOH 2006 pre-congress workshop: Criteria for the case definition of musculoskeletal diseases in the occupational setting, Bologna, Italy, 9 July 2006

Introduction

Musculoskeletal disorders of the upper limb and neck are a common cause of morbidity and lost work time (6, 9, 12). For example, in the UK, data from the Labour Force Survey indicate a self-reported annual incidence of 91 per 1000 adults for work-related illnesses of the upper limb and neck, and an estimated loss of 4.1 million working days per year (5). However, the optimum classification of upper limb disorders (ULDs) remains a cause of debate (1, 7, 11).

Difficulty arises principally because of the multiplicity of disorders, diagnostic labels and different approaches adopted within the field - often, several different names for the same disorder, labels that vary between different clinical specialties, terms like 'RSI', 'cumulative trauma disorder', and 'work-related upper limb disorder' that are ambiguous in terms of coverage and the exact boundaries of definitions, and disagreements about the range of disorders that exist. The relative lack of pathognomic symptoms, signs, and useful investigations, the basic subjectivity of pain reports, and the common uncertainty about the true origin of symptoms all impede attempts to define a suitable gold standard, and thereby hamper the effective pooling of research observations.

One response to this situation has been for panels of experts to agree consensus criteria that can then be exploited in formal investigations or employed in health surveillance. Table 1 lists a few among a substantial number of standardised schemes that have been proposed (10-16).

Table 1 - *Some standardised schemes for classifying disorders of the upper limb*

Research-based:		
Finland	Waris et al (5)	1979
Finland	Viikari-Juntura et al (6)	1983
North America	Silverstein (7)	1985
North America	McCormack et al (8)	1990
Workshop-based:		
England	Harrington et al (9)	1998
	Palmer et al (10)	2000
The Netherlands	Sluiter et al (11)	2001

Recent efforts to address the problems of classification have focused on translating the consensus criteria of 'expert' workshops (proposals with a degree of face and content validity) into workable protocols for use in epidemiological enquiries. In this paper, I discuss some of the criteria by which the success of such schemes could be assessed in the absence of a true gold standard, and describe by way of illustration a scheme developed by our own research group (the Southampton Examination Schedule for Upper Limb Disorders), and the further steps we have taken to evaluate it within the framework I propose. I conclude with a view on the future needs for research planning in this area of inquiry.

Evaluation criteria in the absence of a dependable gold standard

A scheme for use in field epidemiology should be clear, unambiguous and feasible to implement. It should also fulfil basic measurement properties, such as repeatability within- and between-observers, and ideally would show congruence with some other reasonable independent measure of the same end point. More importantly, it should show utility in terms of distinguishing groups that would benefit from different actions - for example, different treatments, different advice on prognosis, or different associations with potentially avoidable risk factors (2).

Three points are worth emphasising. Firstly, a repeatable scheme is not necessarily a correct one, but a non-repeatable one is unlikely to be of value to researchers. Secondly, the main *raison d'etre* for making a diagnosis is to improve decision-making in the management or prevention of illness – if a scheme fails to distinguish groups that could benefit from different actions then the arbitrary labels that are applied carry no 'added value' and the purpose of physical examination becomes questionable (2). Thirdly, the optimal case definition is not necessarily fixed, but may vary according to the study question and study population. (Nonetheless, it would be convenient if a definition could be chosen that had application to several common areas of enquiry and in a variety of study settings.)

The Southampton Examination Schedule for Upper Limb Disorders

For several years the MRC Centre in Southampton has had a programme of work on classification of ULDs in which the logic of development and evaluation has been followed through step by step. Stage 1 comprised a workshop leading to clinical consensus; at stage 2 this was converted into an explicit protocol, and research staff were trained to improve standardisation; at stage 3 we then assessed the schedule's repeatability in various settings and its agreement with an independent reference standard; and more recently, at stage 4, we have evaluated associations with risk factors, disability and treatments received, and prognosis. This paper focuses mainly on the first three stages, although preliminary findings from stage 4 will be mentioned in passing.

Stage 1

The starting point for development of the schedule was a workshop in Birmingham in the UK hosted by the Health and Safety Executive (4). A multi-disciplinary panel of 29 experts were presented with a number of choices for diagnostic criteria. In a so-called Delphi process, they voted, their choices were collated and re-presented for discussion at the workshop, and voted on again. Eventually the process led to consensus criteria for eight different disorders - three at the shoulder, two at the elbow, two at the wrist, with non-specific forearm pain as a diagnosis of exclusion (table 2). Some conditions that appear in schemes proposed by others did not feature in the voting list, and for some like thoracic outlet syndrome there was no agreement.

Stage 2

At the next stage I and my colleagues in Southampton formalised the Harrington criteria in more explicit terms, and filled in some of the gaps with definitions of our own - for example, covering AC joint dysfunction, and subacromial and olecranon bursitis. By this stage the criteria specified the *elements* of a diagnosis, but did not specify the *methods* in detail. A further elaboration was necessary to maximise standardisation and repeatability.

Table 2 - *Diagnostic criteria for upper limb disorders: report of a Delphi consensus workshop (adapted from Harrington et al, 1998 (9))*

Disorder	Diagnostic criteria
Rotator cuff tendinitis	History of pain in the deltoid region AND pain on resisted active movement (abduction - supraspinatus; external rotation - infraspinatus; internal rotation - subscapularis
Bicipital tendinitis	History of anterior shoulder pain AND pain on resisted active flexion or supination of forearm
Shoulder capsulitis (frozen shoulder)	History of pain in the deltoid area AND equal restriction of active and passive glenohumeral movement with capsular pattern (external rotation > abduction > internal rotation)
Lateral epicondylitis	Epicondylar pain AND epicondylar tenderness AND pain on resisted extension of the wrist
Medial epicondylitis	Epicondylar pain AND epicondylar tenderness AND pain on resisted flexion of the wrist
De Quervain's disease of the wrist	Pain over the radial styloid AND tender swelling of first extensor compartment AND EITHER pain reproduced by resisted thumb extension OR positive Finkelstein's test
Tenosynovitis of wrist	Pain on movement localised to the tendon sheaths in the wrist AND reproduction of pain by resisted active movement
Carpal tunnel syndrome	Pain OR paraesthesia OR sensory loss in the median nerve distribution AND ONE OF: Tinel's test positive, Phalen's test positive, nocturnal exacerbation of symptoms, motor loss with wasting of abductor pollicis brevis, abnormal nerve conduction time
Non-specific diffuse forearm pain	Pain in the forearm in the absence of a specific diagnosis or pathology (sometimes includes: loss of function, weakness, cramp, muscle tenderness, allodynia, slowing of fine movements)

Care was taken to specify in as much detail as possible the anatomical landmarks and boundaries, the procedures, manoeuvres, cut points and interpretation. For example, landmarks were defined closely both in words and by line diagrams (figure 1), the procedure for eliciting tenderness or pain on restricted movement was defined, a standardised protocol was laid down for measuring joint movements, and the protocol was supported by a standardised recording *proforma* and a simple training video used to induct new research staff.

Stage 3

As a first step in assessing the schedule's performance, repeatability and validity were assessed in a sample of rheumatology out-patients (8). Replicate examinations were conducted within and between-observers and repeatability estimated in terms of kappa coefficients (for categorical variables) or mean differences and limits of agreement (for measurements of joint movements). We also assessed the agreement between nurse examination and clinic diagnosis, assuming the clinic diagnosis to be an independent reference standard. Altogether, 43 subjects were examined independently, blinded and in random order, by a trained research nurse and a consultant rheumatologist at an interval of a few minutes; 22 patients were also examined twice by the same pre-trained nurse.

Table 3 shows, as an example, the between-observer repeatability of physical signs at the shoulder for the 86 shoulders in 43 people assessed in this way. Agreement by the chance-adjusted kappa coefficient was good to excellent as judged by Fleiss's criteria (3), where a value of 0.40-0.75 is consid-

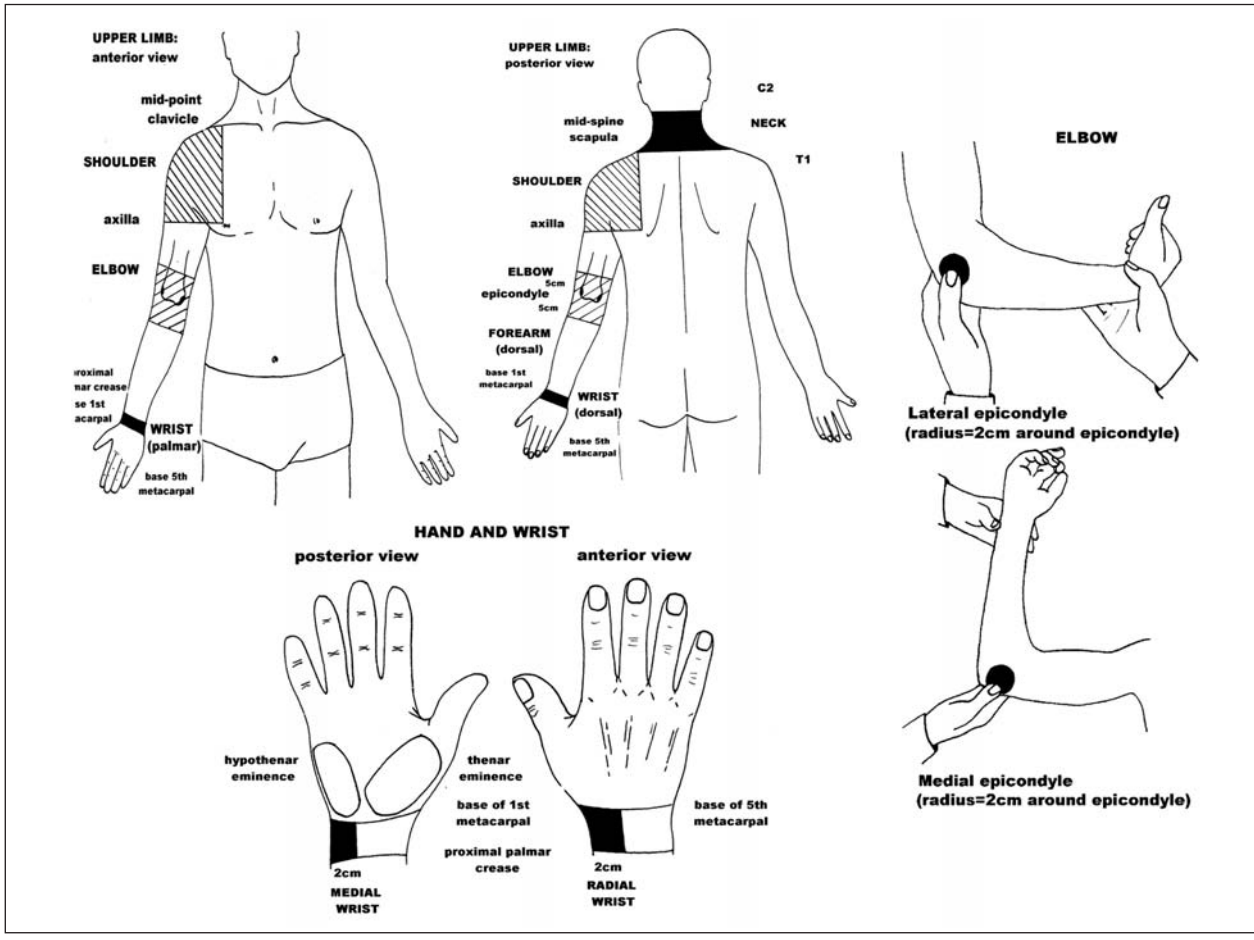

Figure 1 -Standardised line diagrams used in the Southampton Examination Schedule

Table 3 - *Between-observer repeatability of physical signs at the shoulder in the Southampton Examination Schedule (10)*

Signs	No of pairs	Observer1/Observer 2				Kappa coefficient
		-/-	-/+	+/-	+/+	
Tenderness	86	68	2	3	13	0.80
Pain on resisted:						
- elbow flexion	86	70	3	1	12	0.83
- forearm supination	86	73	3	3	7	0.66
- external rotation	86	66	2	1	17	0.90
- internal rotation	86	74	4	3	5	0.54
- abduction	86	67	0	5	14	0.81
Painful arc	86	78	1	0	7	0.93

Observer 1 = nurse
Observer 2 = consultant rheumatologist

ered 'good agreement' and above this is considered 'excellent'. Agreement over measured joint movements was also acceptable, with ≥85% of paired measurements between-observers within 20^0 of one another. Similar agreement was achieved for other sites and measurements.

Table 4 presents the sensitivity and specificity of the nurse assessment in terms of diagnosis, assuming the specialist to be an imperfect but compromise reference standard. In the hospital setting, the schedule had a high specificity (88%-100%) and a reasonable sensitivity (58%-100%).

A more exacting test would be to demonstrate repeatability among the generally milder and more borderline cases expected to exist within a community or workplace sample. We thus nested an enquiry of this kind within a large study of ULD that began with a mailing to over 10,000 working age subjects registered with primary care practices in Southampton (almost everyone in Britain registers with a family doctor for care which is largely free at the point of delivery). Among 6,038 respondents were 3,152 who reported some upper limb or neck symptoms in the past 12 months and were invited to an interview (14). Eventually, 1,960 people (62% of those invited) were examined and classified by the schedule. These included 97 consecutive subjects who were examined twice, independently, by a rheumatologist or nurse in random order (13). In this more challenging setting, the repeatability of physical signs was less good; but agreement over diagnosis was still reasonable with a median kappa of 0.66 and a performance considered 'good' or 'excellent' by Fleiss's criteria for five of the eight disorders assessed (table 4).

The main area of disagreement involved the

Table 4 - *Agreement on diagnosis: Southampton Examination Schedule vs. specialist*

Clinic diagnosis	Nurse vs. Rheumatologist			
	Hospital survey (10)		Community survey (15)	
	Sensitivity (%)	Specificity (%)	Kappa coefficient	
Adhesive capsulitis	87	92	0.39	(0.66)*
Rotator cuff tendinitis	58	88	0.35	(0.40)*
Bicipital tendinitis	100	98	0.49	
Lateral epicondylitis	67	98	0.75	
De Quervain's disease	67	99	0.66	
Carpal tunnel syndrome	67	100	0.93	

* Revised definition (see text)

shoulder, where kappa coefficients for adhesive capsulitis and rotator cuff tendinitis were <0.4. This caused us to refine the protocol. As originally defined, these diagnoses required the patient to identify shoulder pain localised within a particular anatomical distribution, but patients found it hard to be so precise. A post-hoc analysis that relaxed the symptom criteria to allow 'any' shoulder pain improved agreement from 'fair' to 'good'. Even after this accommodation, there still appeared to be a problem in the classification of shoulder disorders, highlighted by data from the whole sample. Among 1,960 subjects, 410 were classified as having adhesive capsulitis; but 205 of these were also counted as having rotator cuff tendinitis, while 35 of 42 subjects with subacromial bursitis were also classified as having adhesive capsulitis (14). It seems therefore that either these conditions co-exist frequently, or the schedule is less discriminating at the shoulder.

Nonetheless, the scheme did subdivide cases into groups with different levels of disability who were treated differently by the medical profession, and this was apparent at the shoulder as well as other sites (table 5) (14). Those counted as having a specific disorder were more likely to report difficulty with activities of daily living, more likely to have received an injection, and more likely to have taken a prescription drug than those with non-specific pain at the same anatomical site. The schedule seemed therefore to track case severity and to display some congruence with clinical practice.

To a degree, it also distinguished groups that had different associations with risk factors. To take two examples from the community survey, the risk factor of arm elevation was associated more strongly at the shoulder with non-specific pain than with a specific diagnosis (odds ratio (OR) 4.9 vs. 1.6), whereas the activity of typing was more strongly associated with a specific hand-wrist disorder than with non-specific hand-wrist pain (OR 3.1 vs. 1.3) (15). Evidence of this kind suggests some added practical value in identifying groups that might benefit from different preventive actions.

Finally, our preliminary findings on subjects from the community survey, followed up at 18 months, suggest that the schedule may have a useful degree of predictive validity. ORs were calculated for persistence of same-site pain in those with a specific diagnosis at the start of follow-up as compared with non-specific pain at the same anatomical site. The odds of persistent shoulder pain were 3.8 times higher (95% confidence interval (CI) 1.1-12.7) in those with a specific shoulder disorder as compared with non-specific shoulder pain, while the OR of persistent hand-wrist pain was 4.7 (95%CI 1.3-17.1) in specific as compared with non-specific hand-wrist cases at baseline (personal communication).

Discussion

Classification of ULDs remains a challenging and contentious area of research enquiry. Candidate schemes are usually developed on the basis of clinical consensus, and thus tend often to incorporate a useful degree of face and content validity. However, it seems desirable to go beyond this and to evaluate the measurement properties of classification schemes (7). In the absence of a true gold standard, evidence of repeatability within- and between-observers in clinical and community settings, and some measure of agreement with an independent and plausible external reference standard seem desirable pre-requisites. More important

Table 5 - *Disability and medical care in the past 12 months: specific vs. non-specific disorders (adapted from reference 14)*

	Shoulder (%)		Elbow (%)		Wrist/hand (%)	
	Specific	Non-specific	Specific	Non-specific	Specific	Non-specific
Impossible to:						
Sleep	2.5	0	2.9	1.1	2.2	1.4
Carry bags	11.5	6.1	20.0	5.6	13.0	7.8
Had an injection	9.0	0	5.7	2.8	6.5	2.2
Took prescribed drugs	37.5	23.2	37.1	26.3	44.6	21.1

in practice though is evidence of utility and predictive validity (2). Despite some limitations, including potential for misclassification at the shoulder, the Southampton Examination Schedule shows useful promise in relation to these criteria, as judged in a variety of ways.

A pressing future goal for research in this area is for investigators to collect data on the component items of diagnosis (symptoms and signs) according to common evidence-based standards. The Southampton Examination Schedule offers one potential route towards this goal. In any event, wider use of a common repeatable standard would facilitate communication between research scientists and allow the effective pooling of data, while the continuing debate about diagnosis could be forwarded by testing empirically among the various choices which definitions 'add value' in terms of furthering the management and/or prevention of ULD cases.

NO POTENTIAL CONFLICT OF INTEREST RELEVANT TO THIS ARTICLE WAS REPORTED

REFERENCES

1. 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*: 141-149
2. COGGON D, MARTYN C, PALMER KT, EVANOFF B: Assessing case definitions in the absence of a diagnostic gold standard. Intl J Epidemiol 2005; *34*: 949-952
3. FLEISS JL: The measurement and control of misclassification error. In Fless JL ed: *Statistical methods for rates and proportions*. Chichester: Wiley, 1981: 140-153
4. HARRINGTON JM, CARTER JT, BIRRELL L, GOMPERTZ D: Surveillance case definitions for work-related upper limb pain syndromes. Occup Environ Med 1998; *55*: 264-271
5. JONES JR, HUXTABLE CS, HODGSON JT, PRICE MJ: *Self-reported work-related illness in 2001/2: Results from a household survey*. Norwich: Health and Safety Executive, 2003
6. MCCORMACK RR JR, INMAN RD, WELLS A, et al: Prevalence of tendinitis and related disorders of the upper extremity in a manufacturing workforce. J Rheumatol 1990; *17*: 958-964
7. PALMER K, COGGON D, COOPER C, DOHERTY M: Work-related upper limb disorders: getting down to specifics. Ann Rheum Dis 1998; *57*: 445-446
8. PALMER K, WALKER-BONE K, LINAKER C, et al: The Southampton examination schedule for the diagnosis of musculoskeletal disorders of the upper limb. Ann Rheum Dis 2000; *59*: 5-11
9. Silverstein BA: *The prevalence of upper extremity cumulative disorders in industry (Thesis)*. The University of Michigan: Occupational Health and Safety, 1985
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*: S3-S102
11. VAN EERD D, BEATON D, COLE D, et al: Classification systems for upper-limb musculoskeletal disorders in workers: a review of the literature. J Clin Epidemiol 2003; *56*: 925-936
12. VIIKARI-JUNTURA E: Neck and upper limb disorders among slaughterhouse workers: an epidemiologic and clinical study. Scand J Work Environ Health 1983; *9*: 283-290
13. WALKER-BONE K, BYNG T, LINAKER C, et al: Reliability of the Southampton examination schedule for the diagnosis of upper limb disorders in the general population. Ann Rheum Dis 2002; *61*: 1103-1106
14. WALKER-BONE K, PALMER KT, READING I, et al: Prevalence and impact of musculoskeletal disorders of the upper limb in the general population. Arth Rheum 2004; *51*: 642-651
15. WALKER-BONE K, READING, I, COGGON D, et al: Risk factors for specific upper limb disorders as compared with non-specific upper limb pain: Assessing the utility of a structured examination schedule. Occup Med 2006; *56*: 243-250
16. WARIS P, KUORINKA I, KURPPA K, et al: Epidemiologic screening of occupational neck and upper limb disorders. Scand J Work Environ Health 1979; *5*: 25-38

ACKNOWLEDGEMENTS: *The community study was supported by a grant from the Health and Safety Executive and a project grant PO552, from the Arthritis Research Campaign. The data were collected by Karen Walker-Bone (ARC Clinical Research Fellowship), Kathy Linaker and Trish Byng. Statistical analysis was by Isabel Reading (Colt Foundation Fellowship). David Coggon and Cyrus Cooper participated in the design, analysis and interpretation of all studies. Infrastructure support was provided by the Medical Research Council. Vanessa Cox and Ken Cox provided support for computer programming; Denise Gould typed this manuscript*