Washington University School of Medicine
Digital Commons@Becker

OHS Faculty Publications | Occupational Health and Safety

2007

A pattern recognition approach to the development of a classification system for upper-limb musculoskeletal disorders of workers

Dorcas E. Beaton
Institute for Work & Health, Toronto, Ontario, Canada

Claire Bombardier
Institute for Work & Health, Toronto, Ontario, Canada

Donald C. Cole
Institute for Work & Health, Toronto, Ontario, Canada

Sheilah Hogg-Johnson
Institute for Work & Health, Toronto, Ontario, Canada

Dwayne Van Eerd
Institute for Work & Health, Toronto, Ontario, Canada

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/ohs_facpubs

Part of the Medicine and Health Sciences Commons

Recommended Citation

Beaton, Dorcas E.; Bombardier, Claire; Cole, Donald C.; Hogg-Johnson, Sheilah; Van Eerd, Dwayne; Evanoff, Bradley A.; and The Clinical Expert Group, "A pattern recognition approach to the development of a classification system for upper-limb musculoskeletal disorders of workers". *Scandinavian Journal of Work, Environment & Health*, 131-139. 2007.

This Article is brought to you for free and open access by the Occupational Health and Safety at Digital Commons@Becker. It has been accepted for inclusion in OHS Faculty Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.

Authors
Dorcas E. Beaton, Claire Bombardier, Donald C. Cole, Sheilah Hogg-Johnson, Dwayne Van Eerd, Bradley A. Evanoff, and The Clinical Expert Group

This article is available at Digital Commons@Becker: https://digitalcommons.wustl.edu/ohs_facpubs/6

Scand J Work Environ Health 2007;33(2):131–139

A pattern recognition approach to the development of a classification system for upper-limb musculoskeletal disorders of workers

by Dorcas E Beaton, PhD, [1–5] Claire Bombardier, MD,[1, 5–7] Donald C Cole, MSc,[1, 8] Sheilah Hogg-Johnson, PhD,[1, 9] Dwayne Van Eerd, MSc,[1, 8] the Clinical Expert Group [10]

Beaton DE, Bombardier C, Cole DC, Hogg-Johnson S, Van Eerd D, the Clinical Expert Group. A pattern recognition approach to the development of a classification system for upper-limb musculoskeletal disorders of workers. *Scand J Work Environ Health* 2007;33(2):131–139.

Objectives Workers' musculoskeletal disorders are often pain-based and elude specific diagnoses; yet diagnosis or classification is the cornerstone to researching and managing these disorders. Clinicians are skilled in pattern recognition and use it in their daily practice. The purpose of this study was to use the clinical reasoning of experienced clinicians to recognize patterns of signs and symptoms and thus create a classification system.

Methods Two hundred and forty-two workers consented to a standardized physical assessment and to completing a questionnaire. Each physical assessment finding was dichotomized (normal versus abnormal), and the results were graphically displayed on body diagrams. At two different workshops, groups of experienced researchers or clinicians were led through an exercise of pattern recognition (clustering and naming of clusters) to arrive at a classification system. Interobserver reliability was assessed (8 observers, 40 workers), and the classification system was revised to improve reliability.

Results The initial classification system had good face validity but low interobserver reliability (kappa <0.3). Revisions were made that resulted in a proposed triaxial classification system. The signs and symptoms axes quantified the areas in the involved upper limbs. The proposed third axis described the likelihood of a specific clinical diagnosis being made and the degree of certainty. The interobserver reliability improved to ~0.70.

Conclusions This triaxial classification system for musculoskeletal disorders is based on clinically observable findings. Further testing and application in other populations is required. This classification system could be useful for both clinicians and epidemiologists.

Key terms nosology, repetitive strain injury, reproducibility of results.

Classification systems, as applied to clinical disorders, are sets of rules that define the minimum criteria to be met to establish the existence of a disorder (1). Meeting (or not meeting) these criteria can be used to determine if a person "has" that disorder. If so, it defines this person as being different from those not meeting the criteria in terms of some relevant aspect of current experience (pathology, pain, disability) or future course (likelihood of a slower or faster recovery, likelihood of response to treatment). Classification systems can, therefore, help

[1] Institute for Work & Health, Toronto, Ontario, Canada.
[2] Department of Occupational Therapy, University of Toronto, Toronto, Ontario, Canada.
[3] Graduate Department of Rehabilitation Sciences, University of Toronto, Toronto, Ontario, Canada.
[4] St Michael's Hospital, Toronto, Ontario, Canada.
[5] Health Care Research Division, The Wellesley Toronto Arthritis and Immune Disorder Research Centre, University Health Network, Toronto, Ontario, Canada.
[6] Clinical Epidemiology and Health Care Research Program, Department of Medicine and Department of Health Administration, University of Toronto, Toronto, Ontario, Canada.
[7] Department of Medicine, Mt Sinai Hospital and University Health Network, Toronto, Ontario, Canada.
[8] Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada.
[9] Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada.
[10] Clinical Expert Group: Peter C Amadio (United States), Rachelle Buchbinder (Australia), Simon Carette (Canada), Bradley A Evanoff (United States), Carol Kennedy (Canada), Christine B Novak (United States), Glenn Pransky (United States), Eira Viikari-Juntura (Finland).

Reprint requests to: Dr DE Beaton, Institute for Work & Health, 481 University Ave., Suite 800, Toronto, Ontario M5G 2E9. [E-mail: dbeaton@iwh.on.ca]

clinicians and researchers communicate about prevalence and incidence rates, the impact of disorders, and prognosis. In order to be effective, they must be both widely accepted and have established value (1–4).

Classification systems often reflect the current theory about the underlying pathophysiology of the disorders in question, and the criteria often reflect this belief in focusing on indicators of the pathology to establish diagnoses (2). However, certain syndromes, particularly those involving primarily symptom-based disorders [such as the chronic fatigue syndrome (5)] or disorders of unknown pathophysiology [such as the Gulf War syndrome (6)], may be difficult to classify in this way. When a disorder cannot be classified according to a specific, clearcut pathology, the possibility exists for several different theories to arise that lead to different criteria and different labels (4, 7). For example, Buchbinder cites 17 different labels for soft tissue pain in the shoulder (8). Such variability in classification may also lead to inconsistencies in studies of burden (4, 9), etiology, and prognosis (10). These inconsistencies can lead to questions about the biological mechanisms involved and even the legitimacy of the disorders (5, 11). Such is the case with the upper-limb musculoskeletal disorders of workers when wide-ranging debates over the causes, pathology, and even existence of these disorders threaten to divert attention away from the real goal of their management—to reduce burden at a personal, workplace, and societal level (12, 13).

Historically, low-back pain went through a similar nosological struggle (14). The resolution was the adoption of a classification system that, in the absence of red flags, abandoned the need to pursue the specific pathology and, rather, described the pattern of pain experience (4, 14). This system, the Quebec taskforce classification of low-back pain, allowed the clinical presentation to drive the classification and has truly facilitated communication in the field. More recently, a similar approach has facilitated communication about whiplash-associated disorders (15, 16) and chronic fatigue syndrome (5). The use of the description of the clinical presentation with or without a clinical diagnosis as the basis of a workplace-based classification system would break with the tradition of purely pathologically based nosologies. However, it would provide a simpler, descriptive system that would have the advantage of describing the entire presentation of upper-limb disorder(s) and provide a framework for communication across different users (epidemiologic case definitions, clinical decision making).

One way to develop such a descriptive system would be to tap into the skills of experienced clinicians and clinical researchers who might recognize patterns of symptoms that are indicative of a more severe condition or a worse prognosis. Clinicians intuitively recognize meaningful patterns of symptoms and signs when they assess a patient. It is their ability to classify this patient as similar or dissimilar to another cluster of patients that they have encountered, or read about, that guides their treatment plan. While this pattern recognition is a skill used in clinical practice, we believe that such skills can also be applied to create a more inductively based classification system for musculoskeletal disorders of the upper limbs.

The purpose of this study was to create a classification system for musculoskeletal disorders of the upper limbs by using the clinical skill of pattern recognition to group workers into clusters based on the similarity and differences in the presentation of the sign and symptoms experienced. This classification system would be for use in workplace studies, but would also be a means for clear communication between research and clinicians. We also sought to establish acceptable levels of interobserver reliability for this system.

Material and methods

Workplace study sample

Our investigation builds on research on the occurrence and burden of musculoskeletal disorders of the upper limbs at a large urban newspaper. An initial cross-sectional survey of 1207 unionized workers was undertaken and produced 1003 usable responses, with findings as reported elsewhere (9, 17). A total of 558 workers agreed to be contacted at a later date for the component of the study reported here. Participants experiencing more intense or frequent episodes, as well as those with episodes of long duration, were over-sampled from the 558 workers willing to be contacted in order to increase the number likely to have relevant physical assessment findings and likely to be in need of classification (see table 1). However, some workers with very mild, transient, or no symptoms were also included in order to provide a full spectrum of findings for classification. Using these criteria, we selected 239 persons for our project who, along with four new workers who had not participated in the first survey, made up a study sample of 243 workers. The sample, although useful for our research question, should not be considered representative of the workplace as a whole. All of the participants underwent a standardized physical assessment, and all but one completed a detailed questionnaire, leaving a sample of 242 workers for the analysis.

Data collection

Each of the 242 workers underwent a standardized 20-minute physical assessment by the same investigator

(DEB). The elements of the examination were determined on the basis of a review of the relevant literature (18) and consensus by a group of clinicians (not the same group of clinicians as participated in this project). It covered the domains of range of motion, muscle strength, pain on resisted motion, sensation, provocative tests, and dolorimetry. The active range of motion was assessed for the neck and also bilaterally for 12 movements in the shoulder, elbow, wrist, and hand. Muscle strength was graded on a 0–5 scale (5 = normal) (18), and pain on resisted motion (5-point scale) was tested in the same planes of motion. As resistance was not used in the assessment of the range of motion of the cervical spine, only pain on active motion was recorded. Sensation was assessed with the use of Semmes-Weinstein monofilament (18) testing of the volar aspect of the index finger, little finger, and the dorsal web space, each representing the autonomous zone for one of the three main peripheral nerves. The provocating tests included the impingement sign, palpation for tenderness over the greater tuberosity, bicipital groove and acromioclavicular joint, Mills' test, Tinel's sign's, Phalen's signs, and the first carpometacarpal joint grind test. Six sites were selected for tender point assessment, which was tested using a dolorimeter with a 1.5-cm^2 rubber end (19, 20). Pressure was applied to the standardized sites and slowly increased to a maximum of 5 kilograms of force. A site was considered tender if pain was experienced at less than 4 kilograms of force (19). A manual available from the present authors describes the methods used in detail. All of the results were dichotomized into abnormal or normal on the basis of available literature (18, 21).

The workers also completed a questionnaire that included a pain diagram. They were asked to carefully shade in the areas in the pain diagram in which they experienced pain, numbness or tingling, or swelling. On an accompanying area of the questionnaire the workers reporting that they had experienced any pain or discomfort in the past year were asked to indicate the affected body region (neck, shoulder, elbow or forearm, or wrist or hand—left or right side). Information about the workers' pain (intensity, frequency, duration) was gathered with the use of a questionnaire. The workers then indicated whether they had experienced pain in the past 7 days and in which regions; this information was considered for current status. They were also asked to rate the level of pain they had experienced in the past 7 days, the average intensity of the pain they had experienced over the past year, and the intensity of the worst pain they had experienced over the past year (22).

The data for each worker was summarized visually on a "worker profile" (figure 1). These body diagrams were coded using colored symbols for each abnormal finding. A detailed legend accompanied this diagram. The workers' pain diagrams were scanned

Table 1. Description of the participants in the various stages of the study leading to the sample used in this analysis (in the far right column).

Case level (three strata)	Phase I participants		Those agreeing to repeat contact		Those recruited for current study (N= 243 getting physical examination)	
	N	%	N	%	N	%
Not a case [a]	455	45.4	206	36.9	36	15.1
Level A [b]	343	34.2	212	38.0	107	44.8
Level B [c]	205	20.4	140	25.1	96	40.2
Total	1003	100	558	100	243 [d]	..

[a] No pain or pain less than required for level A.

[b] Discomfort three times in the past year or lasting more than 5 days (16) and not level B.

[c] Discomfort 12 times in the past year or lasting more than 7 days and of moderate or worse severity (16).

[d] 239 (100%) + 4 new workers.

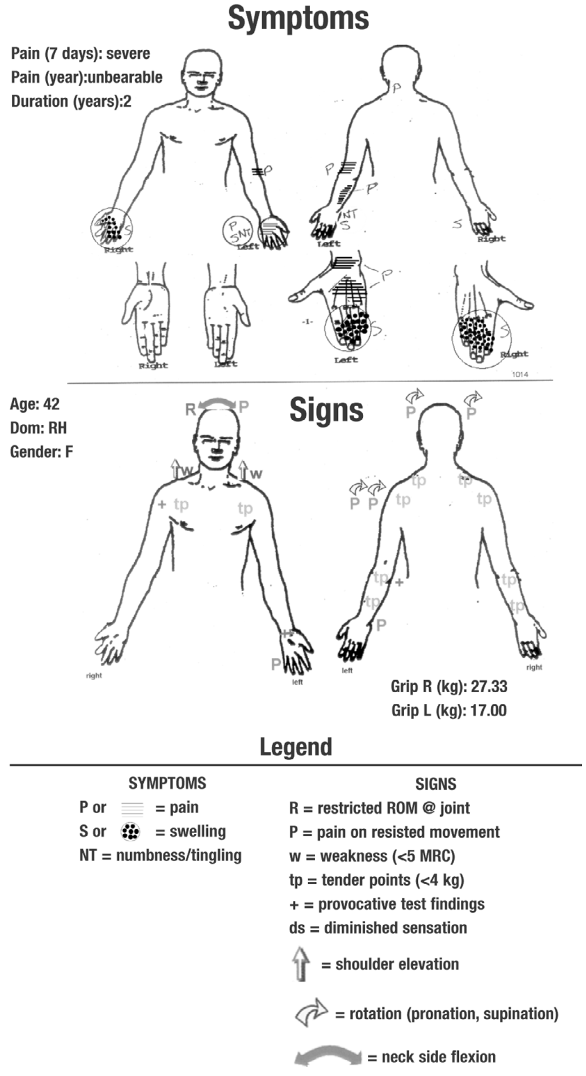

Figure 1. Profile of a worker's symptoms and signs.

into a computer and displayed on the same page as the examination findings for the same worker. Additional text was added beside the pain diagram to describe the worker and some of the symptoms (age, gender, duration of pain, intensity of pain).

Workshops

Three workshops were held, one pilot (for which the data were not used) and two formal, to use pattern recognition to derive a classification system that would sort the worker profiles into meaningful clusters. The workshop participants had experience in epidemiology or the clinical care of persons with musculoskeletal disorders.

The pilot workshop involved five experienced clinicians (some were also researchers) and was held to establish and refine the pattern recognition process. The groupings from this pilot workshop were not used in subsequent analyses, nor did these pilot participants participate in the formal workshops.

Once the pattern recognition process was established, experienced clinicians and epidemiologists were invited to attend one of two formal workshops [Toronto (N=9): all had clinical experience and some were also researchers; Helsinki, Finland (N=19): researchers in the area of work-relevant soft-tissue disorders and most with clinical experience]. In the two formal workshops, the participants were split into two groups (A and B). These groups were sent to separate rooms in which two different sets of 60 randomly selected worker profiles were mounted on the walls. Each workshop participant was given a set of small stickers that were coded as "their" color. They then were asked to place a sticker on any profile that he or she considered to represent a typical or somehow recognizable pattern of symptoms or findings. Considerable discussion took place among the workshop participants throughout. The investigators acted as facilitators and asked the workshop participants, still as two separate groups, to explain why they thought the profiles with stickers were recognizable. This step helped to define what "fit" in that cluster. Following this naming, all of the profiles (including those with no sticker) were then arranged by each group into agreed-upon clusters with similar signs and symptoms. Each cluster was then given a name. The two groups of workshop participants then came together to present their clusters to each other and consolidate the findings into a mutually acceptable set of clusters. The Helsinki workshop groups were not able to consolidate the two lists of cluster labels due to time. Therefore the results of these two formal workshops were three sets of cluster labels describing the patterns of symptoms and signs observed among workers in the workplace study.

Consolidation of findings across workshops

Cluster labels and descriptions from the two formal workshops (along with detailed comments sent by one participant after a workshop) were amalgamated by the investigators (DEB, CB, DCC, SHJ, DVE) into what we have called the consolidated classification system. Similarities and dissimilarities in the three initial sets of cluster labels (two from Helsinki and one from Toronto) were taken into account, as were notes from the workshops themselves. Clusters were consolidated only if they shared themes and structural features and if the resulting consolidated classification system continued to reflect the main issues identified in the source material.

Clinical expert input

After the workshops and consolidation of the findings, eight clinical experts (clinicians or epidemiologists) were assembled as a "clinical expert group". These experts were known to the investigators as having been involved in the classification of upper-extremity musculoskeletal disorders and in the treatment of or research on these problems. Each member agreed to participate in the testing of the interobserver reliability and refinement of the classification system. Professionally there was one orthopedic (hand) surgeon, two physiotherapists, one occupational health physician, one physiatrist, one epidemiologist, and two rheumatologists. Three of the group members had also been members of one of the workshops. The first task of this group was to evaluate the interobserver reliability.

Reliability testing. Forty profiles (as in figure 1) were selected at random from the sample of 242 persons for the evaluation of interobserver reliability. The sample size calculation was based on work by Kraemer & Korner (23) and Donner & Eliasziw (24). The alpha was set at 0.05, the beta at 0.20. $\text{Rho}_{(0)}$ and $\text{rho}_{(1)}$ were set at 0.60 and 0.85, respectively. A sample size of 42.1 was calculated and rounded to 40.

Each clinical expert was sent the consolidated classification system (with detailed instructions) and asked to apply it to the set of 40 randomly selected profiles (the same set of profiles was sent to each expert). The responses were returned, and the agreement between the clinical experts' ratings was assessed using an unweighted kappa statistic.

Revision of the classification system. Two months later, six of the clinical expert group gathered in Toronto for a workshop at which the results of the agreement exercise were presented (the other two were unable to attend). The 242 worker profiles were mounted on the walls of the meeting room so that they could be consulted during the discussions. The group worked through many proposed revisions of the consolidated classification system and set the goal of achieving an interobserver reliability of 0.75. Discussions, facilitated by a small group facilitator, continued over the course of a full day until a consensus was reached on the content, criteria, and operational definitions of a revised classification system.

Reliability testing of the revised classification system. Another 40 profiles were randomly selected and sent to the clinical experts for classification using the revised classification system. The responses were returned, and unweighted kappa coefficients were calculated to see if the changes had improved the reliability.

Results

Workplace study sample

The mean age of the 242 workers was 45 (range 24–65) years. Seventy-one percent was married, 13% was single, and 12% was separated or divorced. Altogether 93% was in full-time permanent employment with the newspaper. The disability scores ranged between 0 (no difficulty in daily tasks) to 57.5/100, with an average of 14.2/100 (SD 12.4) using the Disabilities of the Arm, Shoulder & Hand (DASH) outcome measure (25). General mental and physical health, as measured by the short-form 12-item general health survey (SF-12) were very close to scores obtained with the general population norm 50.

Results of the formal workshops

The formal workshops resulted in clusters in two groups, group A and group B. In Toronto, group A chose to base their clusters on considerations of how well the worker would be expected to do, resulting in cluster labels such as “bad prognosis” and “good prognosis”. The following clusters were the result: good prognosis, peripheral; good prognosis, arthropathy; good prognosis, neurological; good prognosis, neck; bad prognosis, multiple early problems; bad prognosis, rotator cuff; bad prognosis, neurological; bad prognosis, uncertain diagnosis. In contrast, group B focused on the degree of upper-limb involvement (diffuse versus local) and resulted in the following clusters: asymptomatic; local, neck; local, neck + one arm; diffuse. However, groups A and B worked through their differences and arrived at the agreed upon the following list of cluster labels: asymptomatic; local, neurological; local, one site, bilateral; local, one site, unilateral; neck + one arm + neurological findings; neck + one arm; neck, alone; multiple regions or multiple findings or diffuse.

The participants at the Helsinki workshop arrived at two sets of cluster labels but did not have time to consolidate them. The groups were labeled as group A (normal; mild nonspecific; mild cervical; neck tension; neck disorder; possible carpal tunnel syndrome; de Quervains tenosynovitis; neck and shoulder; nonspecific symptoms likely to continue) and group B (no symptoms; significant arm pain; neck pain with radiation; simple neck and trapezius; neck pain and wrist pain; carpal tunnel syndrome; possible de Quervains tenosynovitis; neck and shoulder; complex; arthralgia of proximal interphalangeal joint of the digit. As can be seen, there was some overlap between the groups.

Consolidated classification system

Table 2 shows the first iteration of a consolidated classification system. The investigators decided that the results from the two workshops were best summarized by describing four key features or domains. The first was the zone, referring to the number and location of the major findings. It was viewed as principally a descriptive feature, with many possible combinations. The second feature, the type of disorder, referred to a pattern of signs and symptoms, as well as to the possible presence of an identifiable diagnosis. The third and fourth features were the presence of neurological findings and the duration of symptoms—neither of which had been assessed in the physical assessment and questionnaires as specifically as the workshop participants thought necessary.

Table 2. Consolidated classification—first iteration of the classification system, reflecting a consolidation of the workshop findings.

Domain	Options
Involved zones	Neck, shoulder or upper arm, elbow or forearm, wrist or hand Left and right assessed for at least three regions
Type of disorder	Asymptomatic—no pain or discomfort Diffuse—more than two zones on one or both sides or one zone on each side Regional—one- or two-zone involvement on one side Single specific—one specific disorder on one side* Multiple specific—one or more specific disorders on one or both sides* *Specify disorder(s): ____________________
Neurological signs	Positive neurological signs No neurological signs
Duration of symptoms	1 week or less More than 1 week to 4 weeks More than 4 weeks to 12 weeks More than 12 weeks

Results of work with the clinical expert group

Reliability testing of the consolidated classification system. The kappa statistics for interobserver reliability of the initial iteration of the consolidated classification system were low. Of the four domains of the classification system, "involved zones" had the highest kappa at 0.26. The kappa for "type of disorder" was 0.23. Agreement between the clinical experts was low with regard to the presence of "neurological findings" (kappa 0.17) and a combination of all the axes (kappa 0.09).

Revisions of the classification system. The clinical experts discussed possible reasons for the low reliability coefficients, and each was explored for its potential impact on the poor reliability. During the pattern recognition exercise, the clinical experts reported that neurological status and duration were difficult to quantify with the information made available from the worker profiles, simply because the data were not sufficiently detailed. Instructions for the zones were not clear, and there was some disagreement about the boundaries, mainly the proximal and distal boundaries of the elbow region.

The type of disorder required the clinicians to describe the worker as having one or more specific disorders. Some of the clinical experts were unwilling to make that distinction with only the data provided. They also emphasized the difference between the diagnostic process in clinical practice and the current exercise in deriving diagnoses from cross-sectional survey information. The discussions among the clinical experts led us to understand that epidemiologists tend to gather consistent and more comprehensive data across patients in an attempt to allow criteria for various diagnoses to be applied, whereas clinicians follow a decision map prospectively, only looking at findings along that path until a recognizable cluster emerges. The clinical experts also expressed concern about the lack of a consensus concerning diagnostic criteria for musculoskeletal disorders in the literature. Agreement over diagnostic criteria is essential when a data set such as the current one is being dealt with.

However, the clinical experts decided that there was a need to allow for specific diagnoses in the classification system but that such diagnoses needed to be placed within the context of other findings concerning the entire upper limb. The decision was made to describe the "density", or the number of zones in the upper limb affected in terms of both signs and symptoms. In addition, a separate axis would be available to describe the likelihood that a combination of findings was consistent with one or more specific diagnoses. This final axis requires a set of criteria to define what would be a definite, probable, or possible case of the diagnosis. Such definitions have been derived through consensus work on carpal tunnel syndrome (26), and guidelines for a variety of specific diagnoses have been suggested in study documents of the Joint Programme for Working Life Research in a European Perspective (referred to by the Swedish acronym SALTSA) (27). It was decided that we would not develop this axis in this group, but rather would merge our work, which offers a description of the degree of involvement of the extremities in terms of signs and symptoms, with the work of others such as the SALTSA group outlining diagnostic criteria for specific disorders (26, 27).

During the consensus workshop, the clinical experts decided to eliminate neurological status as a separate category. They felt that some of the specific diagnoses would pick up a neurological finding, as would some of the other examination findings. It was decided, given the difficulty of a thorough neurological examination, to consider any neurologically oriented finding to be a positive indicator of the need for a more in-depth assessment. Neurological findings were therefore incorporated into the description of the signs and symptoms (including their specific and diffuse location in the extremity) as already described.

The revised classification system documents, therefore, the number of regions in the upper limb that had symptoms or signs. Symptoms and signs are on separate axes to allow for symptoms in the absence of signs, and vice versa. The following four levels along each axis were chosen to reflect the degree of involvement of the upper limbs: none, local (one zone only), regional (two zones in one extremity), and diffuse (more than two zones in one extremity or one or more in both extremities). The last level was difficult to define. For example, bilateral pain in thumb extension may be closer to one disorder than right shoulder pain and left thumb extension pain; however, according to the rules, both would be classified as "diffuse".

This revised classification system is shown in table 3. A detailed guide, available from the present authors, includes operational definitions and a body diagram template to allow the assessor to identify the zones involved correctly.

*Reliability of the revised classification system. S*ubstantial improvements in the kappas (0.61–0.73) suggested improved consistency in the application of the revised classification (signs: 0.61, symptoms: 0.73, and overall: 0.65). The levels of interobserver agreement in the revised classification fell into the range that Landis & Koch suggested as "good" reliability (28), and they approximate the level considered acceptable for group level analyses (29). Reliability is a necessary precursor to validity (30), and, with more confidence in reliability, we resolved to continue with this classification system and assess its construct validity.

Discussion

Our paper reports the development of a classification system for musculoskeletal disorders of the upper limbs on the basis of information from 242 workers at a large urban newspaper. The result was a triaxial classification system that describes the degree of involvement of the upper limbs in terms of both symptoms and signs, as well as the likelihood of any specific diagnoses.

The proposed classification system has several strengths. First, it is a system that could prove useful to both the epidemiologist and the clinician. It would meet the needs of epidemiologists because it provides an overall view of the location of pain and discomfort and the location of positive assessment findings. This approach is an improvement over reliance on recorded clinical diagnoses. From the clinician's point of view, it is a useful way of placing their working diagnosis in the context of the patient's entire presentation, which may, in turn, help them better understand the treatment response. The classification system also allows clinicians to describe and retain their level of certainty (possible, probable, definite) around the diagnosis, which could, in turn, aid the epidemiologist in classification decisions. This classification system could, therefore, help facilitate better communication between epidemiologists and clinicians and provide a step that could help bridge the schism currently found among some in these groups (7, 31, 32). Second, it is not the work of only a small group of local researchers (the authors), but has had input from 28 different clinicians and epidemiologists (formal workshop participants) from around the world. As Katz et al (4) suggest, for this type of work to be acceptable, a wide range of clinical perspectives should be brought into the process. Our workshop participants and experts came from clinical backgrounds in medicine and rehabilitation, occupational medicine, and epidemiology. We believe that if we can show agreement between the clinical and epidemiologic fields, the other stakeholders (workers' compensation, industry, labor) will look favorably on the system. Finally, it is a system that is built on the experience of individual workers, as it describes their pattern of symptoms and signs across their upper limbs. In doing so, a classification system like this may be more closely related to how the symptoms manifest themselves in a worker's personal discomfort, health care utilization, and lost productivity—areas of great interest in the management of musculoskeletal disorders. Indeed, the evaluation of the validity of the classification system has also begun and has been reported elsewhere (13). Initial findings suggest that people with more diffuse signs or symptoms have a greater likelihood for lost time, higher pain levels, and higher levels of self-reported difficulty doing their usual work (13).

Table 3. Proposed classification system reflecting revisions made by the investigators and clinical expert group to improve interobserver reliability.

Axis	Options
Symptoms	1. None: asymptomatic, no reported symptoms 2. Local: symptoms in one zone on one side of the body 3. Regional: two zones affected on one side of the body 4. Diffuse: more than two zones in one extremity or one or more zones in both extremities
Signs	1. None: no positive findings in the physical examination 2. Local: findings in one zone on one side 3. Regional: findings in two zones in one extremity 4. Diffuse: more than two zones in one extremity or one or more zones in both extremities
Specific diagnosis [a]	1. None: no symptoms or signs suggestive of a specific diagnosis 2. Possible: symptoms or signs possibly consistent with a specific diagnosis 3. Probable: symptoms or signs probably consistent with a specific diagnosis 4. Definite: symptoms or signs that are consistent with a specific diagnosis

Rather than defining the criteria for the axis of a specific diagnosis, we are seeking the findings of other groups. Since the workshops in Helsinki and Toronto, the SALTSA group (27) has released a compendium of relevant criteria that, should it become accepted, might well fit into the third axis. In the present study, this axis was labeled in accord with the work of the Johns Hopkins group (26), which adopted a scale incorporating possible, probable, and definite diagnoses and allowed for the description of persons who meet some, but not all, of the requisite findings. The result may allow for a common description of the presentation of the worker for the workplace and clinical parties. Clinicians may go further along the "diagnosis" axis to pursue a specific treatment decision and evaluate its effect on that specific pathology, or on the presentation of the worker as a whole. For example, the presentation may be widespread; however, a clinician may choose to pursue the "probable clinical diagnosis" carpal tunnel syndrome. After successful treatment, the whole presentation could be re-evaluated, and note could be made of a reduction in shoulder–neck pain, as well as in hand symptoms.

Interobserver reliability is important for this system to be useful across users. Initially, we had problems with low interobserver reliability, but it improved substantially with a revision of the classification system (more descriptive categories) and improved instructions. The lowest kappa coefficient is now in the "signs" axis, and it is the only one that did not quite make the level considered to be the minimum for group-level analysis (>0.70, 0.75) (33, 34). None made the level of agreement suggested (kappas >0.90) for applying the results to individual patients (that two observers could be confidently assumed to be able to place one specific individual in the same category) (1, 35). Our results indicate that, at the group level, a person could be reasonably confident in the comparability of the two assessors. In clinical or research settings, the observer would likely be dealing directly with the worker or patient and would have access to all levels of data (not just aggregated, abnormal findings as depicted on our diagrams). Reevaluation of the reliability in a primary data collection setting would be helpful. It should be noted, however, that interobserver reliability has repeatedly been shown to be more difficult to achieve than test-retest reliability, a finding we reproduced in a reliability study in preparation for this workplace study (36). We also used a more conservative unweighted kappa coefficient across all observers, rather than a likely higher weighted kappa coefficient.

The limitations of the work must be acknowledged. The developed classification system reflects the experience of a specific group of 242 workers from a large newspaper. It must, therefore, be applied to additional sets of data to determine whether or not it is able to describe the experiences of other groups. The system also requires validation to ensure that it does indeed separate workers into clusters that differ in terms of disease burden or prognosis (4). Finally, the categories depicting the degree of involvement of signs will depend on the physical assessment carried out, and the one used in the present classification may not be considered the best in the opinion of others. In fact, many of the workshop participants and clinical experts suggested physical assessment items that they felt should be included. This issue requires additional investigation, and a standardized minimal examination needs to be defined. Regardless of the fact that some of these “favorite” tests were missing, the workshop participants and clinical experts were able to group the worker profiles according to the clusters defined. Another potential limitation of the proposed triaxial classification system is the great number of potential categories to consider for each worker with symptoms and signs. Perhaps the use of a great number of categories does not resolve the current situation of trying the “fit” a specific diagnosis (from the many available) to a worker. However, viewing the symptoms and signs axes of the systems as descriptive of the complexity of the worker’s state and the potential diagnoses axis as an important axes for directing early and effective treatment parallels the approach of the Quebec task force on acute low-back pain (14) and whiplash-associated disorders (16). Other literature supports the opinion that placing a painful disorder within the context of how widespread the symptoms and signs are has important prognostic value, with the more widespread pain leading to a worse prognosis (37, 38).

Despite these limitations, our paper reports a novel approach to the classification of workers with neck and upper-limb pain or discomfort. As in the experience of the Quebec task force on acute low-back pain (14) and whiplash-associated disorders (16), it may be that, by returning to a simple description of the presentation rather than pursuing very specific diagnoses, we can develop a system that distinguishes patients likely to recover quickly from those who may be slower to recover (2, 4). The often heated debate over the diagnosis and classification of musculoskeletal disorders may, in fact, be hampering the ultimate goal—to advance our understanding of work-related pain and reduce its impact on people’s lives and productivity (4). The classification system described by us, or a future modification of it, may help researchers and clinicians move beyond that debate and allow them to communicate about workers in the same language, and hence advance research efforts and their application.

Acknowledgments

This study was sponsored by a grant from the Center for VDT and Health Research, Johns Hopkins University (99–0003).

Dr Beaton was supported by a PhD fellowship from the Medical Research Council of Canada during the conduct of this study and is now supported by a “New Investigators Award”, Canadian Institutes of Health Research.

The authors would like to thank the workers who were assessed, the *Toronto Star*, the Southern Ontario Newspaper Guild, the RSI Committee at the *Toronto Star*, and the worksite upper-extremity group at the Institute for Work and Health for conducting the studies that led to the data used in our project. We deeply appreciate the work and energy of the workshop participants: the Toronto Pilot, the Toronto Workshop, and the Helsinki Workshop. Special thanks go to the organizers of the 1998 Premus conference in Helsinki, which accommodated our request for space and time to conduct our workshop during their meeting. Finally, the authors acknowledge the assistance of Charmaine Heath and Rhoda Reardon in organizing and facilitating the expert consensus workshop.

References

1. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. Toronto: McGraw-Hill, Inc; 1994.
2. Kassirer JP. Our stubborn quest for diagnostic certainty—a

cause of excessive testing. N Eng J Med. 1989;320:1489–91.

3. Feinstein AR. Unsolved scientific problems in the nosology of clinical medicine. Arch Intern Med. 1988;148:2269–74.
4. Katz JN, Stock SR, Evanoff BA, Rempel D, Moore JS, Franzblau A, et al. Classification criteria and severity assessment in work-associated upper extremity disorders: methods matter. Am J Ind Med. 2000;38:369–72.
5. Bates DW, Buchwald D, Lee J, Kith P, Doolittle TH, Umali P, et al. A comparison of case definitions of chronic fatigue syndrome. Clin Infect Dis. 1994;18 suppl 1:S11–S15.
6. Haley RW, Kurt TL, Hom J. Is there a Gulf war syndrome?: searching for syndromes by factor analysis of symptoms. JAMA. 1997;277:215–22.
7. Van Eerd D, Beaton D, Cole D, Lucas J, Hogg-Johnson S, Bombardier C. Classification systems for upper-limb musculoskeletal disorders in workers: a review of the literature. J Clin Epidemiol. 2003;56(10):925–36.
8. Buchbinder R. The classification of soft tissue disorders of the neck and upper limb for epidemiological research. Toronto: University of Toronto; 1993.
9. Beaton DE, Cole DC, Manno M, Bombardier C, Hogg-Johnson S, Shannon HS. Describing the burden of upper-extremity musculoskeletal disorders in newspaper workers: what difference do case definitions make? J Occup Rehab. 2000;10(1):39–53.
10. Beaton DE. Examining the clinical course of work-related musculoskeletal disorders of the upper extremity using the Ontario Workers' Compensation Board administrative database. Toronto: University of Toronto; 1995.
11. Vender MI, Kasdan ML, Truppa KL. Upper extremity disorders: a literature review to determine work-relatedness. J Hand Surg [Am] 1995;20A(4):534–41.
12. Hadler NM. The semiotics of "upper limb musculoskeletal disorders in workers". J Clin Epidemiol. 2003;56(10):937–9.
13. Beaton D, Bombardier C, Cole DC, Hogg-Johnson S, Van Eerd D, Clinical Expert Group, et al. Reliability and validity of a classification system for upper limb work-related musculoskeletal disorders. Toronto: Institute for Work & Health; 2001. Working paper no 144.
14. Spitzer WO, LeBlanc FE, Dupuis M, Abenhaim L, Belanger AY, Bloch R, et al. Scientific approach to the assessment and management of activity-related spinal disorders: a monograph for clinicians [Report of the Quebec task force on spinal disorders]. Spine. 1987; 12(7S):s4–s55.
15. Freeman MD, Croft AC, Rossignol AM. "Whiplash associated disorders: redefining whiplash and its management" by the Quebec Taskforce: critical evaluation. Spine. 1998;23(9):1043–9.
16. Spitzer WO, Skovron ML, Salmi LR, Cassidy JD, Duranceau J, Suissa S, et al. Scientific monograph of the Quebec Task Force on whiplash-associated disorders: redefining «whiplash» and its management. Spine. 1995;20[8S]:1S–73S.
17. Polanyi M, Cole DC, Beaton DE, Chung J, Wells R, Abdolell M, et al. Upper limb work-related musculoskeletal disorders among newspaper employees: cross-sectional survey results. Am J Ind Med. 1997;32:620–8.
18. Marx RG, Bombardier C, Wright JG. What do we know about the reliability and validity of physical examination tests used to examine the upper extremity? J Hand Ther 1999;24A(1):185–93.
19. Simms RW, Goldenberg DL, Felson DT, Mason JH. Tenderness in 75 anatomical sites. distinguishing fibromyalgia patients from controls. Arthritis Rheum 1988;31(2):182–7.
20. Wolfe F, Smythe HA, Yunus MB, Bennett RM, Bombardier C, Goldenberg DL, et al. The American College of Rheumatology 1990 criteria for the classification of fibromyalgia [Report of the Multicenter Criteria Committee]. Arthritis Rheum 1990;33(2):160–72.
21. Yassi A, Hassard TH, Kopelow MM, Schnabl G. Evaluating medical performance in the diagnosis and treatment of occupational health problems: a standardized patient approach. J Occup Med. 1990;32(7):582–5.
22. Von Korff M, Ormel J, Keefe FJ, Dworkin SF. Grading the severity of chronic pain. Pain. 1992;50(2):133–49.
23. Kraemer HC, Korner AF. Statistical alternatives in assessing reliability, consistency, and individual differences for quantitative measures: application to behavioral measures of neonates. Psychol Bull. 1976;83(5):914–21.
24. Donner A, Eliasziw M. Sample size requirements for reliability studies. Stat Med. 1987;6:441–8.
25. Beaton DE, Davis AM, Hudak P, McConnell S. The DASH (disabilities of the arm, shoulder and hand) outcome measure: what do we know about it now? Br J Hand Ther 2001;6(4):109–18.
26. Rempel D, Evanoff B, Amadio PC, De Krom M, Franklin G, Franzblau A, et al. Consensus criteria for the classification of carpal tunnel syndrome in epidemiologic studies. Am J Public Health. 1998;88(10):1447–51.
27. Sluiter JK, Rest KM, Frings-Dresen MHW. Criteria document for evaluating the work-relatedness of upper-extremity musculoskeletal disorders. Scand J Work Environ Health. 2001;27 suppl 1:1–102.
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
29. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther. 1996;18(5):979–92.
30. Wright JG, Feinstein AR. Improving the reliability of orthopaedic measurements. J Bone Joint Surg [Br]. 1992;74-B(2):287–91.
31. Katz JN, Liang MH. Classification criteria revisited. Arthritis Rheum 1991;34(10):1228–30.
32. Hadler NM. Cumulative trauma disorders: an iatrogenic concept. J Occup Med. 1990;32(1):38–41.
33. Kramer MS, Feinstein AR. Clinical biostatistics, LIV: the biostatistics of concordance. Clin Pharmacol Ther. 1981;29(1):111–23.
34. Nunnally JC. Reliability of measurements. Introduction to psychological measurement. New York (NY): Mcgraw-Hill Book Company; 1970. p 107–31.
35. McHorney CA. Health status assessment methods for adults: past accomplishments and future challenges [review]. Ann Rev Public Health. 1999;20:309–35.
36. Marx RG, Hudak PL, Bombardier C, Graham B, Goldsmith C, Wright JG. The reliability of physical examination for carpal tunnel syndrome. J Hand Surg Br 1999;23B(4):499–502.
37. Murphy KA, Cornish RD. Prediction of chronicity in acute low back pain. Arch Phys Med Rehabil. 1984;65(6):334–37.
38. Crook J, Moldofsky H. The clinical course of musculoskeletal pain in empirically derived groupings of injured workers. Pain. 1996;67(2–3):427–33.

Received for publication: 2 November 2004