Inter-Rater Reliability of Physical Examinations in a Prospective Study of Upper Extremity Musculoskeletal Disorders

Caroline K. Smith, MPH, David K. Bonauto, MD, MPH, Barbara A. Silverstein, PhD, MPH, and Dana Wilcox, PT

Objective: To test the inter-rater reliability of physical examinations for upper extremity musculoskeletal disorders. **Methods:** A total of 111 subjects were examined by both an occupational medicine physician and a physical therapist, who were blinded to each others' tests results and subjects' current symptoms and health histories. **Results:** Inter-rater reliability was moderate to excellent (0.52 to 0.88) for shoulder tests but required the inclusion of symptoms for increased inter-rater reliability of fair to excellent (0.27 to 0.57) for the elbow and hand/wrist tests. **Conclusions:** With the lack of "gold standard" tests, it is important that epidemiologic field studies conduct and report inter-rater reliability testing results between study examiners. It is important for researchers to report the results of inter-rater reliability, so that practitioners can weigh the results of study findings to improve both their diagnosis and treatment of these costly injuries.

Work-related upper extremity musculoskeletal disorders (WRMSDs) are a significant cause of morbidity and lost time in worker populations. Recent estimates of WRMSDs from Washington State indicate that WRMSDs accounted for 27.1% of all State Fund accepted workers' compensation claims and were responsible for 42% of the costs of all claims (costing around $4.1 billion from 1997 to 2005).[1] The Bureau of Labor Statistics, reported that although declining, musculoskeletal disorders still accounted for 29% of all work-related injuries requiring days away from work.[2] Although rates of WRMSDs are decreasing, they still pose a significant burden on employers, workers, and workers' families.

Most epidemiologic studies that focus on WRMSDs include self-reported symptoms as well as physical examinations of the upper extremities. Although costly, it is believed that physical examinations are necessary for a clinical diagnosis of specific upper limb disorders. Although there is no "gold standard" for most upper extremity musculoskeletal disorder examinations, physical examinations allow for clinical diagnoses, whereas self-report symptoms, while accurate, provide only nonspecific indicators of a disorder.[3] Numerous attempts have been made to create consensus for upper limb physical examination protocols,[4–6] although similar tests are used, most differ on how examinations tests are conducted. In large epidemiologic studies, it is often necessary to have more than one physical examiner; in these cases, the inter-rater reliability of examination protocols should be required to determine the repeatability of the clinical examinations and more importantly the reliability of the outcomes measured. Without the existence of "gold standards" for many upper limb musculoskeletal disorders, the reliability of an examination protocol based on clinical data is critical if the results of epidemiologic studies are to be trusted.

Although inter-rater reliability testing has shown adequate to good results in clinical populations for the neck[7] among 52 patients and for rotator cuff lesions among 136 subjects at a teaching hospital,[8] there exists very little inter-rater reliability data for nonclinical populations. The Southampton examination schedule was used on 97 symptomatic subjects from a large population-based survey in the United Kingdom and found generally good inter-rater reliability,[5] and results from a US study of keyboard operators revealed mixed results.[6] The purpose of this study is to evaluate the inter-rater reliability of upper extremity clinical tests by two examiners in a large prospective study of workers in Washington State.

METHODS

As part of a large prospective upper extremity musculoskeletal disorder study, examiners conducted tests of full-time permanent workers in 12 service and manufacturing companies across Western Washington State. Detailed study design and methods are described elsewhere.[9] In brief, sites were chosen if they employed >100 full-time equivalent employees, had sufficient variability in physical exposures, and agreed to let workers participate during normal work hours with no loss of pay or privileges. To be eligible for inclusion in this analysis, subjects had to work full-time (≥30 hr/wk) and could not be a temporary worker. The Washington State Institutional Review Board approved all materials and study protocols.

Subjects

Subjects for the inter-rater reliability tests were randomly selected from 733 subjects in 12 service and manufacturing sites. Subjects completed detailed health history questionnaires and psychosocial surveys and underwent upper extremity physical examinations by trained staff, a board-certified occupational medicine physician, licensed physical therapists, or an occupational health nurse. The only exclusion criterion was the subjects' English proficiency. In an effort to ensure the examination protocol could be replicated, we chose only those subjects who spoke English as a first language, to ensure there would not be a need to stray from protocols to be understood by the examiner. A total of 111 subjects agreed to participate from nine sites where supervisors allowed workers to be off the job for a longer period. Additional informed consent was obtained to have examinations twice in the same day by two examiners; the consent and methods for inter-rater examinations were also approved by the Washington State Institutional Review Board, before enrollment. Examiners were blinded to each others' examination findings, to the self-reported symptoms reported by the subjects, and to medical histories. All subjects received their two examinations on the same day, within 1 hour of each other, in different examination rooms at the worksite. The examiners were randomly selected to be either the first or second examiner.

From the Insurance Services Division (Ms Smith, Dr Bonauto, Dr Silverstein), SHARP Program, Washington State Department of Labor and Industries; and Insurance Services Division (Ms Wilcox), Washington State Department of Labor and Industries, Region 4 Consultation, Olympia, Wash.

Address correspondence to: Caroline K. Smith, MPH, SHARP Program, Washington State Department of Labor and Industries, PO Box 44330, Olympia, WA 98504-4330; E-mail: smcb235@LNI.wa.gov.

DOI: 10.1097/JOM.0b013e3181f4396b

Physical Examination

For the purposes of reliability testing, the two examiners who were responsible for the majority of the physical examinations, the occupational medicine physician, and a physical therapist, were compared. To insure comparability of examination methods across disciplines, the examiners reviewed video of each test they were to perform. Periodically, examiners recalibrated their physical examination methods by double examinations and discussed any disparities between their findings.

The protocol for the standardized physical examinations included visual inspection of the neck and upper extremities for signs of tumors, surgical scars, muscle wasting, swelling, and ganglia. Active range of motion for the head and shoulder abduction were also conducted.

Provocative Tests (Based on the Study by Sluiter et al)[4]

Active Shoulder Abduction

- Painful arc noted in degrees from beginning to ending of pain while in abduction.
- Positive findings were noted when pain began/ended in abduction, degrees were also noted.
- Resisted inward and outward rotation.
- Subjects were asked to rotate arms inward (outward) while examiner resisted movement.
- Any pain in shoulder (epaulette area) was considered a positive finding.

Palpation of Lateral and Medial Epicondyle

- The insertion area of the muscles around the lateral (medial) epicondyle is palpated.
- Any pain in insertion area is recorded and considered a positive examination.

Resisted Wrist Extension With Elbow Extended

- Subject places palms on thighs with elbows extended and hands closed in fists.
- Examiner resists movement when subject was asked to extend wrists.
- Any pain in or about the lateral epicondyle is considered a positive examination.

Resisted Wrist Extension With Elbow at 30 Degrees Flexion

- Subject places his palms on thighs, while elbows are in 30 degrees of flexion and hands closed in fists.
- Examiner resists movement when subject was asked to extend wrists.
- Any pain in wrist extensors is considered a positive examination.

Wrist Flexion and Compression

- With the subject's elbows extended, forearms in pronation, and the wrists in 60 degrees of flexion, examiners apply constant pressure of moderate intensity, with second and fourth fingers over the carpal tunnel for 30 seconds.
- Paresthesiae or numbness in the distribution of the median nerve is recorded at the time of occurrence.
- Any paresthesiae or numbness within the 30 seconds is considered a positive examination.

Finkelstein's Test

- Subject is asked to make a closed fist with thumbs tucked in fingers.
- Examiner deviates the wrists ulnarly.
- Pain at or about the radial styloid is recorded and considered a positive examination.

Statistical Analysis

Results of the physical examinations and presence of symptoms were dichotomized into positive versus negative findings. Reliability was examined using both the percent of agreement and the kappa statistic.[10] The kappa statistic is simply the percent of times that the two examiners agreed with each other on a specific subject's test, adjusted for the proportion of agreements that would be expected by chance. The kappa statistic is the most common measure of "true" agreement for nominal data because it indicates what proportion of agreement exists above what could be expected by chance.[11] Kappa values >0.80 were considered almost perfect, 0.61 to 0.80 excellent, 0.41 to 0.60 moderate agreement, 0.21 to 0.41 fair agreement, 0.0 to 0.20 slight agreement, and <0 no agreement.[12] $\chi 2$ tests were run for descriptive statistics, comparing the inter-rater sample to the remaining subsample. Statistical analyses were performed using Intercooled Stata statistical software, version 8.0 (2003; Stata Corporation, College Station, TX).

RESULTS

General demographic characteristics of the inter-rater testing sample were similar to the larger prospective study sample, with ~50% male and an average age of ~40 years. Minor differences between the samples include that subjects in the inter-rater testing sample were more likely to be high school graduates (88.3% vs 83.5%) and self-report being white (65.8% vs 59.5%) than in the larger prospective study (Table 1). Thus, we expect that the inter-rater sample is quite representative of the entire cohort, which, in turn, is likely to be representative of a population doing similar work.

Inter-Rater Reliability of Physical Examination Tests

Data were available for 91 subjects for shoulder abduction and inward and outward rotation of the shoulders and 111 were available for the remaining tests. Shoulder abduction tests were not

TABLE 1. Demographic Characteristics of the Inter-Rater Sample (n = 111) Compared With Entire Prospective Cohort (n = 733)

	Inter-Rater Sample		Entire Cohort		
	n	%	*n*	%	*P**
Male	56	50.5	383	52.3	0.68
Age (yr)					0.31
≤40	51	46.0	370	50.5	
>40	60	54.1	363	49.5	
Race					0.042
White	73	65.8	436	59.5	
Asian/Pacific islander	10	9.0	132	18.0	
Other	28	25.2	165	22.5	
Hispanic ethnicity	14	12.6	95	13.0	0.31
High school graduate	98	88.3	612	83.5	0.17

*Based on χ^2 comparing the inter-rater sample to the balance of subjects within the entire cohort.

TABLE 2. Percent Agreement and Kappa of Physical Examination Tests by Examiner

Test	Number of Matched Pairs	Physician/Physical Therapist*				Observed	Expected	Kappa	95% CI†
		+/+	−/−	+/−	−/+				
Active shoulder abduction	91								
Right side		8	76	5	2	92.3	77.9	0.65	0.38, 0.83
Left side		5	79	5	2	92.3	83	0.55	0.25, 0.77
Resisted shoulder abduction	91								
Right side		5	82	4	0	95.6	85.7	0.69	0.36, 0.88
Left side		3	83	5	0	94.5	88.49	0.52	0.18, 0.78
Resisted outward rotation of shoulder	91								
Right side		5	82	1	3	95.6	85.77	0.69	0.36, 0.88
Left side		2	88	1	0	98.9	94.65	0.79	0.28, 0.96
Resisted inward rotation of shoulder	91								
Right side		4	86	0	1	98.9	90.59	0.88	0.49, 0.98
Left side		2	86	3	0	96.7	92.55	0.56	0.16, 0.84
Resisted wrist extension with the elbow extended	111								
Right side		7	97	5	2	93.69	82.83	0.63	0.36, 0.82
Left side		5	97	7	2	91.89	84.25	0.49	0.21, 0.71
Palpation of the lateral epicondyle	111								
Right side		6	68	11	26	66.67	64.69	0.05	−0.12, 0.24
Left side		4	76	12	19	72.07	72.07	0.04	−0.11, 0.25
Palpation of the medial epicondyle	111								
Right side		8	74	15	14	73.87	67.67	0.19	0.005, 0.40
Left side		8	75	13	15	74.77	68.2	0.21	0.02, 0.41
Resisted wrist extension with the elbow in 30 degree of flexion	111								
Right side		4	105	1	1	98.2	91.4	0.79	0.41, 0.94
Left side		1	109	1	0	99.1	97.33	0.66	0.13, 0.94
Finkelstein's test	111								
Right side		4	90	6	11	84.68	79.91	0.24	0.03, 0.48
Left side		2	94	6	9	86.49	84.31	0.14	−0.03, 0.42
Wrist flexion and compression	111								
Right side		2	103	2	4	94.59	91.38	0.37	0.09, 0.69
Left side		4	103	2	2	96.4	89.77	0.65	0.30, 0.86

*+, positive finding, −, negative finding. All results are comparing physician with physical therapist.
†95% CI, 95% confidence interval for the kappa statistic.

performed on 20 subjects, either they declined or time limitations precluded a complete examination. In general, percent agreement was very good, ranging from 66.7 (palpation of right medial epicondyle) to 98.9 (resisted inward rotation of right shoulder) (Table 2). Kappa statistics for 9 of the 20 tests were considered excellent (0.61 to 0.80) and 4 of the 20 were considered moderate (Table 2). The lowest kappa statistics were for palpation of the lateral epicondyle (0.24 right to 0.04 left) and medial epicondyle (0.19 right to 0.21 left) as well as for the Finkelstein's test (0.24 right to 0.14 left).

For inter-rater subjects who reported in their health interview to having symptoms (mild, moderate, severe, or very severe), n = 52 to 73 depending on the body part tested, Kappa statistics were calculated (Table 3). Both observed percent agreement and kappa statistics fell slightly for all shoulder tests but rose dramatically for palpation of the lateral and medial epicondyle and Finkelstein's test (Table 3). When the sample was restricted to those who reported severe or very severe symptoms, both percent agreement and kappa statistics increased for all tests, data not shown, as it was only available for 7 of the 20 possible tests. Palpation of the lateral epicondyle (right side) showed increased percent agreement, from 78.7 for any symptoms to 86.1 for severe/very severe symptoms, and kappa values increased as well from 0.32 to 0.47 (data not shown). Similar increases were seen for palpation of the medial epicondyle tests (bilaterally), with kappa statistics rising from 0.53 to 0.65 for the right side and from 0.42 to 0.46 on the left, although confidence intervals for all tests remained quite wide.

DISCUSSION

Overall inter-rater reliability was good (showing fair to excellent agreement) for the physical examination tests of this working cohort, with the exception of tests from the whole inter-rater sample for lateral and medial epicondyle and Finkelstein's tests for the hand/wrists. When only subjects with symptoms were considered, kappa statistics were similar for most tests in the whole sample, but it dramatically improved for both palpation of the lateral and medial epicondyle and Finkelstein's tests. Kappa scores improved again when those with severe or very severe symptoms were considered (data not shown), although there were too few pairs for all but seven of the body side tests and confidence intervals remained wide. A larger study sample size is needed to further test

TABLE 3. Percent Agreement and Kappa of Physical Examination Tests by Examiner for Subjects With Any Self-Reported Symptoms

Test	Number of Pairs	Physician/Physical Therapist*				Percent Agreement		Kappa	95% CI†
		+/+	−/−	+/−	−/+	Observed	Expected		
Active shoulder abduction									
Right side	33	7	20	4	2	88.5	69.3	0.63	0.34, 0.82
Left side	29	4	20	3	2	89.5	78.4	0.51	0.18, 0.76
Resisted shoulder abduction									
Right side	33	5	24	4	0	93.4	79.5	0.68	0.33, 0.87
Left side	29	3	22	4	0	91.2	82.2	0.51	0.15, 0.77
Resisted outward rotation of shoulder									
Right side	33	5	24	1	3	93.4	79.6	0.68	0.33, 0.87
Left side	29	2	26	1	0	98.3	91.6	0.79	0.27, 0.96
Resisted inward rotation of shoulder									
Right side	33	4	28	0	1	98.4	86.3	0.88	0.48, 0.98
Left side	29	25	2	2	0	94.7	88.3	0.55	0.14, 0.84
Resisted wrist extension with the elbow extended									
Right side	32	5	21	5	1	88.5	73.6	0.57	0.26, 0.78
Left side	23	4	14	3	2	86.5	72.5	0.51	0.19, 0.75
Palpation of the lateral epicondyle									
Right side	32	4	19	2	7	78.7	68.7	0.32	0.04, 0.56
Left side	23	4	9	3	7	76.9	68.3	0.27	−0.01, 0.54
Palpation of the medial epicondyle									
Right side	32	3	23	4	2	86.9	72.4	0.53	0.22, 0.75
Left side	23	3	13	5	2	80.8	66.6	0.42	0.13, 0.67
Resisted wrist extension with the elbow in 30 degree of flexion									
Right side	51	1	49	1	0	97.3	87.2	0.79	0.40, 0.94
Left side	71	0	70	1	0	98.1	94.5	0.66	0.12, 0.94
Finkelstein's test									
Right side	51	0	40	3	8	91.8	82.7	0.53	0.20, 0.77
Left side	71	1	60	3	7	90.6	87.7	0.24	−0.01, 0.64
Wrist flexion and compression									
Right side	51	1	47	1	2	95.9	93.4	0.38	0.05, 0.77
Left side	71	2	67	1	1	96.2	89.3	0.65	0.19, 0.90

*+, positive finding; −, negative finding. All results are comparing physician with physical therapist.
†95% CI, 95% confidence interval for the kappa statistic.

the change in kappa coefficient by increase in symptom severity. The lack of any statistical differences in demographics between the inter-rater testing sample and the larger pool of subjects suggests that these 111 subjects were quite similar to the entire cohort, which is in turn, likely to be representative of workers in similar industries. Thus, the physical examinations examination results for the entire cohort may be generalized to a larger population carrying out similar work. Although inter-rater or interexaminer testing provides critical information when evaluating evidence from epidemiological field studies where physical examinations are included, data on inter-rater reliability are not often published. In addition to the importance of inter-rater reliability is the detailed description of examination tests and protocols because there are no "gold standard" tests for most upper extremity musculoskeletal disorders. Sluiter et al[4] and more recently the Southampton examination schedule,[5] which is based on consensus criteria from the Birmingham Workshop,[13] have attempted to standardize both the examination tests and protocols for use in the general population. Walker-Bone et al found generally good inter-rater reliability for the Southampton examination schedule in a general population. Although we have briefly reported aggregated inter-rater reliability for lateral epicondylitis,[14] we appear to be the first to test the reliability of many of the tests in the examination protocol of Sluiter et al[4].

Similar to Kryger et al[15] who performed inter-rater reliability in computer workers, we found increasing occurrence of physical findings with increasing pain scores, although we did not have a large enough sample to model our results as they did. Similar to Salerno et al.,[6] we had relatively low kappa coefficients for inter-rater reliability among the hand/wrist tests (Table 2); they believed these low numbers to be due to low prevalence of disorders in their population, which might have also been true for our study, although reliability increased markedly for palpation of the lateral and medial epicondyle and Finkelstein's tests in our study, when we limited the results to those with hand wrist symptoms in the past 7 days of examination (Table 3).

Of note, we did find kappa values for tests where the examiner resisted the subject's active motion were generally higher than those tests relying on the examiner applying force on the subject (eg, palpation of the epicondyle, Finkelstein's test, and

carpal compression test). Although speculative, there may be reasons for this observation. First, when subjects initiate movement, they do so in relationship to their ability to exert maximal effort, while the examiner titrates the resistance to that effort. If both examiners can exceed the subject's effort level through resistance of the subject's motion, there is more consistency in the application of the test and therefore higher inter-rater agreement. Second, despite efforts to insure comparable performance of the examination between the two examiners, slight variation in technique and the force levels applied likely exist. Differences in anthropometry across subjects (eg, fat near the epicondyle) might augment differences in force applications by the examiners yielding poorer kappa. Epidemiologic studies using multiple examiners should work toward eliminating differences in examination techniques and examination performance between examiners. Assessing the contribution of variation between examiners, with and without review of the standardized protocol for physical assessment of WRMSDs, may lead to better understanding of variation in prevalence estimates.

In the absence of a valid and reliable "gold standard" for most upper extremity musculoskeletal disorder physical examination tests, it is important that epidemiologic field studies report both the physical examination tests/protocols and any inter-rater reliability data, in addition to study findings, so that practitioners and researchers alike may weigh this information along with the study findings. There exists good evidence-based data on physical examination tests and protocols, which should be used whenever possible, in future epidemiologic studies, and the inter-rater reliability should be reported, preferably stratified by pain intensity or presence of symptoms, to assist practitioners with the diagnosis and treatment of upper extremity musculoskeletal disorders.

ACKNOWLEDGMENTS

This study was supported in part by grant OH 07316 from the National Institute for Occupational Safety and Health (NIOSH) and additional funding from the Washington State Department of Labor and Industries, Olympia, WA.

REFERENCES

1. Silverstein B, Adams D. Work-related Musculoskeletal Disorders of the Neck, Back and Upper Extremity in Washington State, 1996–2004. Olympia, WA: SHARP Program, Washington State Department of Labor and Industries; 2007.
2. Bureau of Labor Statistics, Labor. Nonfatal Occupational Injuries and Illnesses Requiring Days Away from Work, 2007. United States Department of Labor, Publication USDL 08-1716. Washington, DC: Bureau of Labor Statistics; 2008.
3. Walker-Bone KE, Palmer KT, Reading I, Cooper C. Criteria for assessing pain and nonarticular soft-tissue rheumatic disorders of the neck and upper limb. *Semin Arthritis Rheum.* 2003;33:168–184.
4. Sluiter JK, Rest KM, Frings-Dresen MH. Criteria document for evaluation of the work-relatedness of upper extremity musculoskeletal disorders. *Scand J Work Environ Health.* 2001;27:1–102.
5. Walker-Bone K, Byng P, 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.
6. Salerno DF, Franzblau A, Werner RA, et al. Reliability of physical examination of the upper extremity among keyboard operators. *Am J Ind Med.* 2000;37:423–430.
7. Viikari-Juntura E. Interexaminer reliability of observations in physical examinations of the neck. *Phys Ther.* 1987;67:1526–1532.
8. Ostor AJ, Richards CA, Prevost AT, Hazleman BL, Speed CA. Interrater reproducibility of clinical tests for rotator cuff lesions. *Ann Rheum Dis.* 2004;63:1288–1292.
9. Silverstein B, Viikari-Juntura E, Fan ZJ, Bonauto D, Bao S, Smith C. Natural course of nontraumatic rotator cuff tendinitis and shoulder symptoms in a working population. *Scand J Work Environ Health.* 2006;32:99–108.
10. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20:37–46.
11. Sim J, Wright C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther.* 2005;85:257–268.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.
13. 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.
14. Fan ZJ, Silverstein BA, Bao S, et al. Quantitative exposure-response relations between physical workload and prevalence of lateral epicondylitis in a working population. *Am J Ind Med.* 2009;52:479–490.
15. Kryger AI, Lassen CF, Andersen JH. The role of physical examinations in studies of musculoskeletal disorders of the elbow. *Occup Environ Med.* 2007;64:776–781.