AMERICAN JOURNAL OF INDUSTRIAL MEDICINE 37:423–430 (2000)

Reliability of Physical Examination of the Upper Extremity Among Keyboard Operators

Deborah F. Salerno, PhD,[1] Alfred Franzblau, MD,[1,2]* Robert A. Werner, MD,[1,2,3] Kevin C. Chung, MD, MS,[4] J. Steven Schultz, MD,[5] Mark P. Becker, PhD,[6] and Thomas J. Armstrong, PhD,[2,7]

Background *Physical examination is a traditional outcome measure in epidemiological research. Its value as a reliable measure depends, in part, on the prevalence of positive findings. The purpose of this paper is to determine the empirical reliability of physical examination and anthropometry in a field study of upper extremity disorders among keyboard operators.*

Methods *Two experienced examiners independently performed common provocative tests and procedures in physical examinations of the neck and upper extremity among 160 keyboard operators. Two additional examiners conducted anthropometric surveys among 137 workers. Inter-examiner reliability was assessed with observed agreement, kappa statistics, and intra-class correlations (ICC).*

Results *Observed agreement was between 96% and 100% for neck and upper extremity signs, muscle stretch reflexes, and muscle strength, however, with the exception of provocative tests, reliability statistics were unstable. Among the provocative tests, Phalen and Tinel tests had modest agreement after adjusting for chance (κ range: 0.20–0.43). The carpal compression test had the best reliability ($\kappa = 0.60$ and $\kappa = 0.67$, left and right side, respectively). The ICCs for anthropometry ranged from 0.36–0.91.*

Conclusions *Results from the study showed that statistically, except for the carpal compression test, physical examination contributed minimal reliable information. This was attributed mainly to the low prevalence of positive findings, and generally mild nature of upper extremity disorders in this population. The results are the best estimate of what would be found in a field study with experienced examiners. While it may reduce bias, separating physical examination from medical history may contribute to the poor reliability of findings. With a shift toward reliable measures, resources can be allocated to more effective tools, like questionnaires, in epidemiological research of upper extremity disorders among keyboard operators.* Am. J. Ind. Med. 37:423–430, 2000.

KEY WORDS: carpal tunnel syndrome; occupational medicine; physical examination; reliability; upper extremity disorders

[1]Department of Environmental Health Science, School of Public Health, The University of Michigan, Ann Arbor, MI
[2]Center for Ergonomics, School of Engineering, The University of Michigan, Ann Arbor, MI
[3]Physical Medicine and Rehabilitation, Veterans Administration Hospital, Ann Arbor, MI
[4]Section of Plastic and Reconstructive Surgery, Department of Surgery, The University of Michigan Medical Center, Ann Arbor, MI
[5]Department of Physical Medicine and Rehabilitation, School of Medicine, The University of Michigan, Ann Arbor, MI
[6]Department of Biostatistics, School of Public Health, The University of Michigan, Ann Arbor, MI
[7]Department of Industrial and Operations Engineering, School of Engineering, The University of Michigan, Ann Arbor, MI
*Correspondence to: Alfred Franzblau, Department of Environmental and Industrial Health, The University of Michigan, School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: afranz@umich.edu

Accepted 18 October 1999

INTRODUCTION

Physical examination is a basic diagnostic tool, and a typical outcome measure in medical research [Gelberman et al., 1983; Golding et al., 1986; Novak et al., 1992; Toomingas et al., 1999; Feuerstein et al., 1999; Homan et al., 1999]. As demands grow for health screenings and examination, valid and reliable research methods are needed to substantiate results.

Although validity and reliability are important issues, little evidence exists on the reliability of physical examination for upper extremity disorders [Marx et al., 1999]. Clinical evaluation has found fair-to-good reliability of physical examination of the neck between two examiners for most findings among a group of 52 patients [Viikari-Juntura, 1987]. Reliability among 12 patients with suspected carpal tunnel syndrome (CTS) among six examiners (two occupational health workers, two hand surgeons, and two hand therapists) showed mixed reliability among 7 tests [Marx et al., 1998]. In rating tendon reflexes, reliability at best, was fair, among groups of two or three physicians out of 37 physicians [Manschot et al., 1998]. Acceptable reliability was reported for the two-point discrimination test between two examiners among 30 patients [Dellon, Mackinnon, and Crosby, 1987]. ''Spectrum bias'' [Ransohoff and Feinstein, 1978] has been implicated as a reason for divergent results of common clinical tests [Gerr and Letz, 1998].

In contrast to clinical research, this study was conducted among keyboard operators with emphasis on the utility of physical examination in field studies. In this report, we evaluate the reliability of physical examination of the upper extremity, describe certain drawbacks in physical examination as currently performed, and identify key statistical issues underlying study designs. Also, reliability is assessed for anthropometric variables (e.g., wrist width and depth), due to their importance as covariates of nerve function [Stetson et al., 1992; Pierre-Jerome et al., 1997; Salerno et al., 1998].

The fact that this study involved active workers is the key, since the measures were tested in a ''real'' work environment, rather than a clinic requiring extrapolation to industrial medicine. As such, this study provides a more relevant assessment of the utility of physical examination in that it directly examines the value of physical examination in epidemiological field research.

METHODS

As part of a large 2-stage medical survey, examiners conducted tests of keyboard operators at a data coding center in the midwestern United States. In Stage 1, the survey was comprised of a physical examination of the participants, an anthropometric survey, nerve conduction studies, a self-administered upper extremity questionnaire, and a functional activity questionnaire. Three weeks later, workers returned for a slightly modified survey in Stage 2. All participants provided informed consent that had been approved by the University of Michigan Human Subjects Review Committee.

Physical Examination

Inter-examiner reliability of physical examination was assessed for two examiners who were board certified in physical medicine and rehabilitation (JSS), and general surgery with fellowship training in hand surgery (KCC). Both were members of the faculty at the University of Michigan Medical Center.

In Stage 1, an independent standardized physical examination was performed twice on each subject, once by each examiner. The protocol included visual inspection of the neck and upper extremities for signs of muscle wasting, swelling, tenderness, redness, warmth, scars, deformity, nodules, and ganglia. Active range of motion (ROM) was assessed for neck flexion (0–45 degrees), extension (0–45 degrees), rotation (− 45–45 degrees), and lateral bending (45 degrees to the left and right). The shoulders were assessed for active abduction and adduction, resisted abduction and adduction, resisted internal and external rotation, resisted flexion and extension. Assessment was carried out for pain on palpation over the bicipital tendon, resisted elbow flexion and extension, and resisted forearm pronation and supination. Pain, crepitus, and limited ROM were assessed for the elbow and wrist during active ROM. Pain in the hand, dorsal wrist, forearm (volar and dorsal aspects), and lateral and medial elbow was assessed during resisted wrist extension and flexion, respectively. Pain was also assessed in the lateral and medial elbow, and volar and dorsal aspects of the forearm on resisted forearm pronation and supination, and finger flexion and extension. Locking or clicking was assessed on repeated finger flexion. Bilateral biceps, brachioradialis, and triceps muscle stretch reflexes were tested and scored [Hallett, 1993], as also were bilateral muscle strength in the biceps, triceps, deltoids, and opponens pollicis [Guarantors of Brain, 1994].

Four provocative tests were performed.

1. *Finkelstein maneuver* [Hoppenfeld, 1976] was performed by stabilizing the forearm, and instructing the subject to make a fist with thumb tucked inside other fingers. Pain proximal to the thumb on ulnar deviation indicated strong evidence of stenosing tenosynovitis, and was scored as ''positive'' if there was no discomfort when the maneuver was repeated with the thumb extended.

2. *Phalen wrist flexion test* [Phalen, 1966] was performed by maintaining maximal voluntary wrist flexion for a period of one minute. The test was considered positive if symptoms were elicited in the distribution of the median nerve.
3. *Tinel percussion test* [Tinel, 1915] was performed with percussion over the palmar aspect of the wrist. The test was considered positive if the subject reported tingling or pain in the distribution of the median nerve.
4. *Carpal compression test* [Durkan, 1991] was performed by applying pressure manually with two thumbs directly over the flexor retinaculum. The test was considered positive if subjects reported numbness, tingling, or dysesthesia in the distribution of the median nerve within one minute (since both Phalen and carpal compression tests increase intracarpal canal pressure, the same duration of one minute was used).

A static two-point discrimination test was conducted with an esthesiometer to determine tactile sensibility, and was considered abnormal if subjects were unable to perceive two points separated by a 4-mm difference over the fingerpad of digit II. (For logistical reasons, the two-point discrimination test was conducted by examiners in the anthropometric survey, described below.)

The protocol was reviewed and practiced by examiners prior to data collection to maximize conformity. In total, the two examiners independently rated workers on 275 items during the physical examination, which was limited to the neck, shoulders, and upper extremities. Examiners were masked to results of other evaluations in the survey. In particular, examiners were masked to symptoms and medical history data, which were collected separately on self-administered questionnaires. There was no physical examination in Stage 2 of the survey.

Anthropometric Survey

Inter-examiner reliability of anthropometry was assessed for two examiners, one of whom was trained in occupational medicine, and the other was board certified in internal medicine and occupational medicine.

In Stage 1, the protocol included measurement of height, weight, finger circumference (digit II proximal phalanx), finger length (digit II metacarpal-phalangeal joint to the tip), wrist width and depth at the distal crease, and right triceps skinfold thickness. Finger circumference and length was measured with a tape measure. Wrist dimensions and skinfold thickness were measured with calipers (Country Technology, Gays Mills, WI).

In Stage 2, the protocol was the same as in Stage 1 except that neither height nor weight was measured, hence reliability for height and weight was not assessed.

Statistical Analysis

The results of all items on physical examination were categorized according to positive findings on provocative testing. For the dichotomous data, reliability was assessed by two methods: (1) the overall observed agreement, and (2) kappa, a measure of agreement corrected for chance. The kappa statistic [Cohen, 1960] is defined as:

$$\kappa = (p_{\text{Observed}} - p_{\text{Expected}})/(1 - p_{\text{Expected}}),$$

where p_{Observed} is the observed proportion of agreement, and p_{Expected} is the expected proportion of agreement. Weighted kappa statistics (quadratic weights) were used to assess ratings for muscle stretch reflexes and muscle strength. Values of kappa >0.75 were considered excellent; values between 0.40–0.75 were fair to good; and values <0.40 represented poor agreement beyond chance [Fleiss, 1981].

Inter-examiner reliability of the anthropometric survey was assessed with the intraclass correlation coefficient (ICC) as a measure of agreement. The ICC combines a measure of correlation with a test in the difference of means [Kramer and Feinstein, 1981]. Pearson product-moment correlations were used as measures of association, as observations may disagree sharply, yet still be correlated [Müller and Büttner, 1994]. In addition, paired *t*-tests were used to see whether, overall, examiners had the same mean measurements.

Statistical analyses were performed using Stata Statistical Software: Release 5.0 [Stata Corp, 1997].

RESULTS

Of the 161 participants in Stage 1, 138 (86%) participants returned in Stage 2 to complete the survey. Study participants ranged in age from 20–58 years. The average age of participants was 35 years; most were female (91%) and right-handed (91%). Average work-tenure was 1.4 years (0.4 years) with a range from 0.4–1.6 years. All participants had graduated from high school, and two-thirds had formal education beyond high school. There were no significant demographic differences between the participants who completed Stage 1 only and those who went on to complete Stage 2.

Inter-examiner Reliability of Physical Examination

Data from 160 subjects were analyzed (one subject did not participate in both examinations) for evaluation of the reliability of physical examination. Observed agreement was between 96% and 100% for upper extremity signs of muscle wasting, swelling, tenderness, redness, warmth, scars, deformity, nodules, and ganglia; and signs of

TABLE I. Upper Extremity Maneuvers and Provocative Tests Among Midwestern Keyboard Operators

Maneuvers and tests	Positive findings		Observed % agreement	Expected % agreement	Kappa	95% CI	P_1[a]	P_2[a]	*P* value[b]
	Examiner 1	Examiner 2							
Finkelstein maneuver									
Right (n = 158)	7	4	94	93	0.15	(0, 0.30)	0.04	0.03	
Left (n = 158)	6	2	97	95	0.49	(0.36, 0.62)	0.04	0.01	0.05
Phalen test									
Right (n = 151)	18	29	85	73	0.43	(0.27, 0.58)	0.12	0.19	0.02
Left (n = 151)	18	30	84	73	0.41	(0.26, 0.56)	0.12	0.20	0.01
Tinel test									
Right (n = 159)	5	6	96	93	0.34	(0.19, 0.50)	0.03	0.04	
Left (n = 159)	3	6	96	94	0.20	(0.06, 0.35)	0.02	0.04	
Carpal compression test									
Right (n = 159)	29	31	90	69	0.67	(0.52, 0.83)	0.18	0.20	
Left (n = 159)	24	33	88	70	0.60	(0.44, 0.75)	0.15	0.21	0.04
Phalen, Tinel or CCT[c]									
Right (n = 160)	32	40	86	65	0.61	(0.45, 0.76)	0.20	0.25	
Left (n = 160)	27	43	85	65	0.57	(0.42, 0.72)	0.17	0.27	<0.01

[a] P_1 = Prevalence of condition reported by Examiner 1; P_2 = Prevalence of condition reported by Examiner 2.
[b] Unless noted, differences in the prevalence of conditions reported by Examiner 1 and Examiner 2 are not statistically significant at $\alpha = 0.05$ level with the McNemar χ^2 test for independent proportions.
[c] CCT = Carpal compression test.

inflammatory arthritis (swelling, tenderness, redness, and/or warmth), and degenerative arthritis (bony deformities) in the wrist, metacarpophalangeal, and interphalangeal joints. After accounting for chance agreement however, inter-examiner reliability measured by kappa, was low. This was attributed to the low prevalence of positive signs (at most, 7%). Examiner 1 reported more signs than Examiner 2, but the absolute magnitudes of differences were generally small.

Likewise, for range of motion, there was very low prevalence (less than 5%) for most tests with few exceptions: for neck extension, Examiner 1 reported none while, Examiner 2 reported 14 subjects with limited ROM; for pain in the neck on lateral bending, Examiner 1 reported 9 subjects with positive findings, while, Examiner 2 reported none.

Similar to the upper extremity signs, the observed agreement for upper extremity muscle stretch reflexes was excellent (99% agreement), but due to low variance of findings, kappa indicated a low level of agreement after correcting for chance. Both examiners rated all reflexes as normal.

There was also excellent agreement in rating muscle strength. Both examiners rated bilateral biceps and triceps strength as normal power (grade 5). For the deltoids and opponens pollicis, except for ratings in the left thumb of one subject and right deltoids of three subjects that Examiner 1 rated as grade 4, both examiners rated muscle strength as normal power.

In general, there was excellent agreement for the provocative maneuvers and tests, with examiners demonstrating 84%–97% agreement (Table I). But again, kappa values were low, which was mainly attributed to the low prevalence of abnormalities (e.g., the Finkelstein maneuver and Tinel test had at most, 4% prevalence). The kappa value was highest for the carpal compression test ($\kappa = 0.67$, right side). There was higher prevalence of abnormal results on the Phalen test (between 12%–20%) and carpal compression test (between 15%–21%). For the combined results of Phalen, Tinel, and carpal compression tests, kappa values were 0.61 and 0.57, right and left sides, respectively. The low prevalence of abnormal findings on the two-point discrimination test produced unstable reliability, measured by the kappa statistic.

Inter-examiner Reliability of Anthropometric Survey

In the anthropometric survey, analyses of reliability were limited by those subjects who were measured in both Stages 1 and 2 (n = 137). Except for wrist ratio, all ICC values were in the good to excellent range (Table II). The finger circumference had the best reliability (ICC = 0.91).

DISCUSSION

The reliability of physical examination in epidemiological research depends, in part, on statistical issues. For

TABLE II. Inter-examiner Reliability of Anthropometry Among Midwestern Keyboard Operators

Measurement (n = 137)	Examiner 1 Mean (SD)[Range]	Examiner 2 Mean (SD) [Range]	Pearson correlation	Paired *t*-test (*P* value)	Intra-class correlation
Finger circumference (mm)					
Right (n = 136)	66 (5) [55–82]	65 (5) [54–79]	0.93	<0.01	0.91
Left	65 (5) [55–81]	64 (6) [53–80]	0.92	<0.01	0.88
Finger length (mm)					
Right	88 (5) [72–104]	90 (6) [79–112]	0.88	<0.01	0.78
Left	88 (5) [75–104]	90 (5) [79–110]	0.90	<0.01	0.85
Wrist width (mm)					
Right	53 (4) [47–65]	56 (4) [48–70]	0.89	<0.01	0.76
Left	53 (4) [45–66]	56 (4) [47–69]	0.89	<0.01	0.69
Wrist depth (mm)					
Right	37 (3) [32–47]	39 (3) [33–49]	0.80	<0.01	0.63
Left	37 (3) [31–48]	39 (3) [33–50]	0.80	<0.01	0.65
Wrist ratio (depth/width)					
Right	0.70 (0.04) [0.60–0.78]	0.70 (0.03) [0.62–0.80]	0.39	0.01	0.36
Left	0.70 (0.03) [0.61–0.80]	0.70 (0.03) [0.62–0.78]	0.43	0.41	0.43
Right triceps skinfold (mm)	(n = 113)	(n = 113)			
Average[a]	34 (12) [5–62]	43 (12) [14–65]	0.82	<0.01	0.75

[a]Summary statistics for average skinfold do not include subjects (n = 24) who were off the scale in Stage 1 or Stage 2.

instance, the power to generate reliable results presupposes a sizeable prevalence of positive findings. Initially our goal was to test the reliability of physical examination among keyboard workers, yet a discussion on reliability would be remiss without emphasizing that we are answering another, more practical question that is related to the value of physical examination in epidemiological field research.

Notable concerns have been found in reviews of reliability [Koran, 1975a, 1975b], with fair to good inter-examiner agreement of physical signs (κ range: 0.51–0.74). Typically, the studies under review had a small number of subjects, but a variable number of examiners (range: 2–10), unlike the present study with a large number of subjects and two examiners.

Observed agreement between examiners in the present study was between 96–100% for upper extremity signs, muscle stretch reflexes, and muscle strength. However, kappa statistics, accounting for chance agreement, were generally unstable.

The instability of kappa was attributed to the low prevalence of positive findings in this population. When prevalence is very low (i.e., < 10%), the value of kappa approaches zero, and likewise when prevalence is very high (i.e., > 90%) [Thompson and Walter, 1988]. For instance, signs of upper extremity conditions were reported at most among 7% of the study cohort.

Not surprisingly, in the study among patients referred for neurosurgical evaluation [Viikari-Juntura, 1987], the prevalence of positive findings was higher (between 12–20%) than in the present study among active workers. Most kappa values were between 0.24–0.56 for tenderness to palpation in the neck and shoulder. Kappa values for tests related to muscle atrophy and strength ranged from 0.32–0.81 with most prevalences above 10%. Even among a patient group with more severe disease status than active workers, low prevalence of positive findings precluded calculation of some kappa statistics. A standardized protocol and patient cooperation were cited as important factors affecting reliability in the study. In the present study, given the low prevalence of positive findings in this worker population, greater standardization of tests may not have substantially improved the reliability of physical examination.

One concern, particularly with the upper extremity tests, is that participants learned a response due to prompting from the first examiner. This issue was not addressed in the present study since the examination sequence was not recorded, however, there was no known systematic bias in the order of examination. Viikari-Juntura [1987] did test whether a recent examination had any effect on the findings of the second examination by having two raters examine patients in mixed order, and found that although there was often more pain on the second test, the difference was not statistically significant. A related concern is that of median nerve irritation, from one test to the next. However, Marx et al. [1999] found that the order in which tests were conducted was not significant. Only one of their 12 patients, who had a negative Phalen test when first examined, reported a positive test at the time of reexamination.

Common Upper Extremity Tests

Among the provocative tests, the carpal compression test ($\kappa = 0.67$ and $\kappa = 0.60$, right and left hands, respectively) had the best reliability, indicating that it may be the test of choice for evaluating CTS. Phalen and Tinel tests, showed overall high agreement, but poor reliability (κ range: 0.20–0.43).

Variability in the Phalen, Tinel, and carpal compression tests may be attributed, in part, to the nature of these tests. The Phalen test relies on wrist flexion to increase intracarpal canal pressure, the Tinel test relies on mechanical deformation of the nerve, while the carpal compression test uses direct contact pressure to raise intracarpal canal pressure. Increasing intracarpal canal pressure can create local ischemia resulting in a conduction block associated with numbness and tingling in the distribution of the median nerve. Direct compression more effectively raises intracarpal canal pressure, whereas the flexion in the Phalen test causes only mild elevation in intracarpal canal pressure [Werner, Bir, and Armstrong, 1994]. The success of these tests depends on the technical acumen of the examiner as well as the ability of the examiner to engage patient cooperation, and the physical condition of the wrist, particularly with the Phalen test where pain or pathology may prevent full flexion [Paley and McMurtry, 1985].

Inter-examiner Reliability of Anthropometry

In the anthropometric survey, inter-examiner reliability was rated generally good to excellent, as expected (fingers and wrists were not likely to change in dimension during the 3-week study interval). Wrist width and depth had good reliability (ICC range: 0.63–0.76) in comparison to the wrist ratio (depth/width) (ICC = 0.36 and ICC = 0.43, right and left ratio, respectively). One rationale for the poor reliability of the ratio is that the variance of a ratio may greatly exceed the variances of individual components. Another explanation is that there was actually poor agreement between examiners for the wrist ratio.

While most paired *t*-test results were statistically significant, none was practically important. For example, finger circumference was statistically different between examiners (Table II), but comparison of means revealed only a 1 mm difference. This was attributed to the power of the test to detect small differences given the large sample size, and illustrates the importance of tempering statistical significance with practicality.

Practical Implications

Performing a good physical examination takes time, and is an important part of medical practice. Like most clinician-based methods, physical examination has an important role in clinical settings, and in epidemiological studies of populations with a high prevalence of disorders. However, the low prevalence and generally mild nature of upper extremity disorders in certain work settings could justify elimination of physical examination from epidemiological study protocols, with concomitant time and cost savings. Other methods in occupational research, such as questionnaires, are more expedient and reliable than physical examination, but questionnaires may underestimate problems in the neck, elbows, and hands. This was found in a study among 165 women employed in repetitive industrial or varied work tasks [Ohlsson et al., 1994]. Similar to the present study, the protocol involved masking of data collected from questionnaires. Unlike the present study, the examiner solicited symptom data, and if symptoms were reported, a detailed examination was performed in the relevant region. While the severity of symptoms could have differed, 94 (57%) of the women reported neck or upper extremity symptoms on the questionnaire, whereas 140 (85%) of the workers had physical findings, and 75 (45%) were given clinical diagnoses. The researchers concluded that detailed physical examination was imperative for accurate assessment of upper extremity disorders.

In contrast, the present study had few findings on physical examination, yet 61% of participants in the medical survey reported symptoms on the self-administered questionnaire. The striking difference in physical examination findings between work populations (i.e., keyboard operators vs. workers in industrial or varied work tasks) illustrates the limit of generalizability among active workers, and the importance of symptom data. In fact, self-administered measures of upper extremity conditions, such as questionnaires, may be more reliable than physical examination in a population of active workers [Franzblau et al., 1997].

Medical History

In an effort to reduce the subjective element of examination, the physical examination protocol prohibited examiners from soliciting medical history in this study, which effectively restricted a critical piece of clinical information. This feature may have adversely affected examiners' capacity to perform reliably since medical history is an important part of clinical examination, although knowledge of medical history may bias the results [Elmore et al., 1997].

Strengths and Limitations

In any reliability assessment of examiners, their training and experience is crucial. This study benefited from seasoned examiners, including a hand surgeon and a

physiatrist. Thus, results represent a best-case scenario with highly trained, experienced examiners. Reliability may be worse among clinicians who do not possess similar training and experience. Also, the study design reflected actual clinical practice with two examiners performing independent evaluations among a series of subjects.

Paradoxically, the strength of this study limits its generalizability. The study is subject to potential bias due to the number of examiners ($n = 2$) in the assessments. While results may not be representative of large samples from a population of clinicians, findings may be best estimates of what can be found among clinicians trained in a standardized methodology.

The most significant limitation is the low prevalence of abnormalities on physical examination. This finding limits the generalizability of results on reliability to all clinical settings, but nevertheless, has important implications for epidemiological field research of upper extremity disorders among keyboard operators. If repeated study corroborates these findings among keyboard operators, especially among those with longer tenure in the industry, other, more valuable methods, such as questionnaires, should be employed to quantify morbidity in the study of keyboard workers.

CONCLUSIONS

Reliability affects the degree of confidence we have in key results. Results from this study showed that observed agreement was high, yet when accounting for chance, inter-examiner reliability of physical examination was poor or modest for the majority of tests. These findings are the best estimate of what would be found among highly trained, experienced examiners. Among the provocative tests, the carpal compression test was most reliable, suggesting it may be the best test for evaluating CTS among workers. Inter-examiner reliability of anthropometry was rated generally good to excellent.

The key point was that low prevalence of positive findings on physical examination led to unstable reliability. Given the low prevalence, statistically, it is unlikely that greater standardization of techniques would increase reliability. This presents a serious challenge to the use of standard physical examination in epidemiological studies among keyboard workers. For this reason, elimination or reduction of physical examination could be justified, with the realization that an important feature of the present study design was the separation of physical examination and medical history review. A tailored examination in concert with a review of medical history may be needed for reliable results, although this may bias the examination results.

Physical examination is an essential component of clinical medicine, and an important part of establishing and maintaining the physician-patient relationship. However, physical examination, as currently performed, has not been shown to be an effective screening procedure for use in occupational research with largely asymptomatic or mildly symptomatic keyboard operators. Suitable measures, such as questionnaires, designed and tested specifically for active workers, have been shown to be more reliable in such settings.

ACKNOWLEDGMENTS

This study was conducted with financial support by Johns Hopkins University Center for VDT and Health Research, and other sources. The authors gratefully acknowledge the workers and management who participated in the study, and express appreciation to Mike Gerard, Mari Hagen, Deborah Heany, I-Wei Huang, Wendi Latko, and Randy Rabourn for assistance in data collection.

REFERENCES

Cohen J. 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46.

Dellon AL, Mackinnon SE, Crosby PM. 1987. Reliability of two-point discrimination measurements. J Hand Surg 12A:693–696.

Durkan JA. 1991. A new diagnostic test for carpal tunnel syndrome. J Bone and Jt Surg 73-A, no.4:535–538.

Elmore JG, Wells CK, Howard DH, Feinstein AR. 1997. The impact of clinical history on mammographic interpretation. JAMA 277, no.1: 49–52.

Feuerstein M, Burrell LM, Miller VI, Lincoln A, Huang GD, Berger R. 1999. Clinical management of carpal tunnel syndrome: A 12-year review of outcomes. Am J Ind Med 35:232–245.

Fleiss JL. 1981. Statistical Methods for Rates and Proportions. 2nd ed. New York: John Wiley and Sons.

Franzblau A, Salerno DF, Armstrong TJ, Werner RA. 1997. Test-retest reliability of an upper extremity discomfort questionnaire in an industrial population. Scand J Work, Environ and Health 23:299–307.

Gelberman RH, Szabo RM, Williamson RV, Dimick MP. 1983. Sensibility testing in peripheral-nerve compression syndromes. J Bone and Jt Surg 65-A, no.5:632–638.

Gerr F, Letz R. 1998. The sensitivity and specificity of tests for carpal tunnel syndrome vary with the comparison subjects. J Hand Surg 23B, no.2:151–155.

Golding DN, Rose DM, Selvarajah K. 1986. Clinical tests for carpal tunnel syndrome: An evaluation. Br J Rheumatol 25:388–390.

Guarantors of Brain. 1994. Aids to the Examination of the Peripheral Nervous System. London: Baillière Tindall.

Hallett M. 1993. NINDS Myotatic Reflex Scale. Neurology 43:2723.

Homan MM, Franzblau A, Werner RA, Albers JA, Armstrong TJ, Bromberg MB. 1999. Agreement between symptom surveys, physical examination procedures and electrodiagnostic findings for the carpal tunnel syndrome. Scand J Work Environ Health 25, no.2:115–124.

Hoppenfeld S. 1976. Physical Examination of the Spine and Extremities. Norwalk, CT: Appleton–Century–Crofts.

Koran LM. 1975a. The reliability of clinical methods, data, and judgments. NEJM 293:642–646.

Koran LM. 1975b. The reliability of clinical methods, data, and judgments (Second of two parts). NEJM 293:695–701.

Kramer MS, Feinstein AR. 1981. Clinical biostatistics LIV. The biostatistics of concordance. Clin Pharmacol Ther 29, no.1:111–123.

Manschot S, van Passel L, Buskens E, Algra A, van Gijn J. 1998. Mayo and NINDS scales for the assessment of tendon reflexes: between observer agreement and implications for communication. J Neurol Neurosurg Psychiatry 64:253–255.

Marx RG, Bombardier C, Wright JG. 1999. What do we know about the reliability and validity of physical examination tests used to examine the upper extremity? J Hand Surg 24A:185–193.

Marx RG, Hudak PL, Bombardier C, Graham B, Goldsmith C, Wright JG. 1998. The reliability of physical examination for carpal tunnel syndrome. J Hand Surg 23B, no.4:499–502.

Müller R, Büttner P. 1994. A critical discussion of intraclass correlation coefficients. Stat Med 13:2465–2476.

Novak CB, MacKinnon SE, Brownlee R, Kelly L. 1992. Provocative sensory testing in carpal tunnel syndrome. J Hand Surg 17B: 204–208.

Ohlsson K, Attewell RG, Johnsson R, Ahlm A, Skerfving S. 1994. An assessment of neck and upper extremity disorders by questionnaire and clinical examination. Ergonomics 37, no.5:891–897.

Paley D, McMurtry RY. 1985. Median nerve compression test in carpal tunnel syndrome diagnosis: Reproduces signs and symptoms in affected wrist. Orthopaedic Review XIV, no.7:41–45.

Phalen GS. 1966. The Carpal-Tunnel Syndrome: Seventeen years' experience in diagnosis and treatment of six hundred fifty-four hands. J Bone and Joint Surg 48A, no.2:211–228.

Pierre-Jerome C, Bekkelund SI, Mellgren SI, Nordstrom R. 1997. Quantitative MRI and electrophysiology of preoperative carpal tunnel syndrome in a female population. Ergonomics 40, no.6:642–649.

Ransohoff DF, Feinstein AR. 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. NEJM 299:926–930.

Salerno DF, Franzblau A, Armstrong TJ, Werner RA, Albers JW, Bromberg MB. 1998. Nerve conduction studies among workers: Normative values. Muscle Nerve 21:999–1005.

Stata Statistical Software: Release 5.0 College Station, TX. Stata Corporation. 1997.

Stetson DS, Silverstein BA, Wolfe RA, Albers JW. 1992. Effects of age, sex, and anthropometric factors on nerve conduction measures. Muscle and Nerve 15:1095–1104.

Thompson WD, Walter SD. 1988. A reappraisal of the kappa coefficient. J Clin Epid 41, no.10:949–958.

Tinel J. 1915. Le signe du ''fourmillement'' dans les lésions des nerfs périphériques. La Presse Médicale 47:388–389.

Toomingas A, Nilsson T, Hgberg M, Lundström R. 1999. Predictive aspects of the abduction external rotation test among male industrial and office workers. Am J Ind Med 35:32–42.

Viikari-Juntura E. 1987. Interexaminer reliability of observations in physical examinations of the neck. Physical Therapy 67:1526–1532.

Weinstein MC, Fineberg HV. 1980. Clinical Decision Analysis. Philadelphia: WB Saunders Company.

Werner RA, Bir C, Armstrong TJ. 1994. Reverse Phalen's maneuver as an aid in diagnosing carpal tunnel syndrome. Arch PM and R 75, no.7:783–786.