SPINE Volume 28, Number 19, pp 2222–2231

Reliability of Clinical Tests in the Assessment of Patients With Neck/Shoulder Problems—Impact of History

Bo C. Bertilson, MD,* Marie Grunnesjö, DN,* and Lars-Erik Strender, MD, PhD†

Study Design. A clinical trial on patients receiving neck/shoulder physical examinations.

Objectives. To analyze reliability of clinical tests, prevalence of positive findings in the assessment of neck/shoulder problems in primary care patients, and the impact of history, including pain drawing, on these parameters.

Summary of Background Data. Reliability of clinical tests varies, perhaps partly because of the impact of history. To our knowledge, this has not been studied before.

Methods. Two examiners independently assessed 100 patients with a set of 66 clinical tests divided into 9 categories. Half of the patients were examined with and the other half without knowledge of history. Reliability as expressed by percentage agreement, kappa coefficients, and prevalence of positive findings was calculated.

Results. Reliability of clinical tests was poor or fair in several categories and did not alter with history. Only a bimanual sensitivity test reached good kappa values. With known history, prevalence of positive findings increased. Bias was apparent in all test categories except sensitivity tests. Four out of five patients were diagnosed to have neurogenic dysfunction in the affected area.

Conclusions. Our sensitivity test was the most reliable and also exempt from bias and should be studied further. Some common tests may not be reliable. History had no impact on reliability of our tests but increased the prevalence of positive findings. Neurogenic dysfunction seems very common in patients with neck and/or shoulder problems and should be screened for. [Key words: neck/shoulder pain, physical examination, reliability, impact of history, neurogenic dysfunction] **Spine 2003;28:2222–2231**

The diagnostic procedure is the basis of medical decision-making. History and physical examination are considered the most important sources of patient information.[1,2] The physical examination usually follows history to confirm suspected pathology. Thus, history sets the tone and practically determines the form of the physical examination, thereby influencing the diagnostic procedure itself. Could it be that history also influences the reliability of clinical tests and/or the prevalence of positive findings in the physical examination?

Studies of patients with back pain show that there is a relatively weak agreement between the results of physical examination and history.[3] Leclaire *et al*[4] found that the diagnostic accuracy of even experienced clinicians was less than chance when history and the physical examination were assessed on simulators of back pain. Vroomen *et al*[5] found that consistency in overall diagnosis increased kappa values from 0.40 to 0.66 when history was added to the physical examination in assessing lumbar nerve root involvement. We have not found any report on how history may influence the reliability of a specific clinical test in the examination of the upper spine. Patients with neck and/or shoulder problems are of special interest as their symptoms are interpreted to have a predominately psychosocioeconomic origin[6] with few anticipated (and noted) clinical findings. Yet our experience is that careful neurologic testing reveals neurogenic dysfunction, that is, clinical signs of neurologic disturbed sensory and/or motor function in the affected and expected area or organ. The reliability and validity of our clinical tests need to be evaluated.

This study is part of an investigation of patients with neck and/or shoulder problems designed to evaluate the reliability of pain drawing, history, and clinical tests and their respective possibilities for predicting/finding neurogenic dysfunction. The aim of this paper is to discuss reliability of clinical tests, prevalence of positive findings in the assessment of primary care patients with neck and/or shoulder problems, and the impact of history on these parameters.

Materials and Methods

Examiners and Patients. Between November 1998 and April 1999, a physician (examiner B) and a doctor of Naprapathy (examiner M)—a manual therapist certified in the Swedish health care system— both working at Torvalla (a clinic specializing in pain diagnostics and sports medicine) examined 100 patients included in the study. Patients were recruited by having the six primary-care units in Haninge, south of Stockholm, consecutively refer all who met the inclusion criteria: neck and/or shoulder problem with or without radiating pain, age 16 to 66 years, speaks and understands Swedish. Exclusion criteria were: previous examination at Torvalla for neck and/or shoulder problems in the last 3 years or factors contraindicating a complete examination, such as serious infection or cancer. Referral could be done with or without prior examination and/or treatment. Referred patients were informed by our secretary by phone and letter about the study and asked if they

From the *Torvalla Sports Medical Clinic, Stockholm, and the †Division of Family Medicine Stockholm, Department of Clinical Sciences, Karolinska Institutet, Stockholm, Sweden.
Acknowledgment date: January 3, 2002. First revision date: June 4, 2002. Second revision date: October 31, 2002. Acceptance date: January 24, 2003.
The manuscript submitted does not contain information about medical device(s)/drug(s).
No funds were received in support of this work. No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this manuscript.
Address correspondence and reprint requests to Bo C. Bertilson, MD, Family Medicine Stockholm, Alfred Nobels allé 12, Huddinge 141 83, Sweden; E-mail: bo.bertilson@klinvet.ki.se

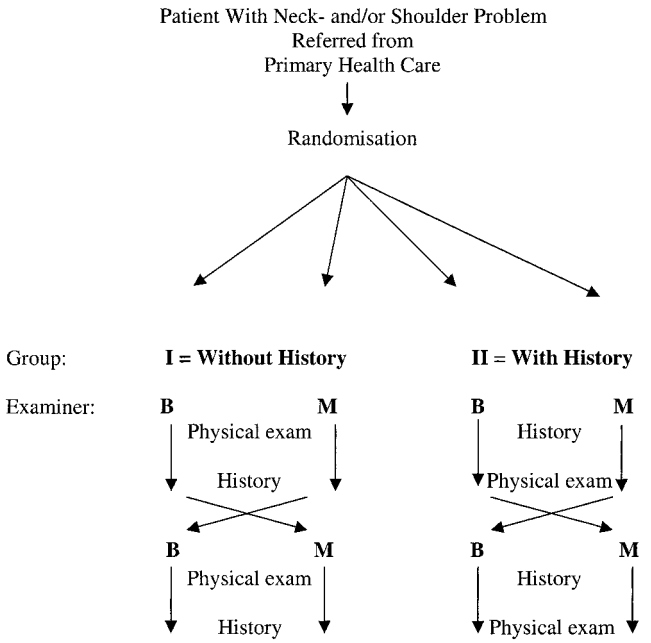

Figure 1. Study flow.

wanted to participate. No one declined. A total of three patients were not enrolled because they did not meet the language criteria. One examiner's assessment of one patient was lost; therefore, the interexaminer figures are calculated on 50 + 49 patients. Approval was obtained from the regional ethics committee.

Randomization and Procedure. A secretary scheduled participating patients into pairs at the same appointment. The first of the two patients to arrive was assigned to one of the two examiners in one of two groups (Figure 1) according to a randomization list made up by an independent statistician. The second patient in the pair was listed, and not randomized, in the same group but started with the other examiner. The list was balanced at every eighth pair and prepared for 160 patients in case of dropouts. Examiners had no insight as to how patients were listed. Before seeing an examiner, each patient filled out a pain drawing (Figure 2) and a history questionnaire that were copied and carried by the patient in an envelope. They were then introduced to the allotted examiner.

In Group I, the physical examination was performed without knowledge of history. Only questions as to whether a certain clinical test was positive could be asked. The findings of the clinical tests were noted on a standardized protocol. Afterwards, the history was made known to the examiner, but no change in the physical examination protocol could be made. Patients would then change rooms, and the protocols were handed to the secretary. The other examiner would, without knowledge of the previous examination, repeat the procedure.

In Group II, the examiner would first look at the pain drawing and the history questionnaire and then interview the patient. The physical examination was then performed, and the protocols were handed to the secretary before patients changed rooms and let the second examiner repeat the procedure.

During assessment sessions, patients were not told any results and were instructed not to discuss or reveal anything from previous sessions. Afterwards, patients were informed by the physician about findings, given recommendations and, when

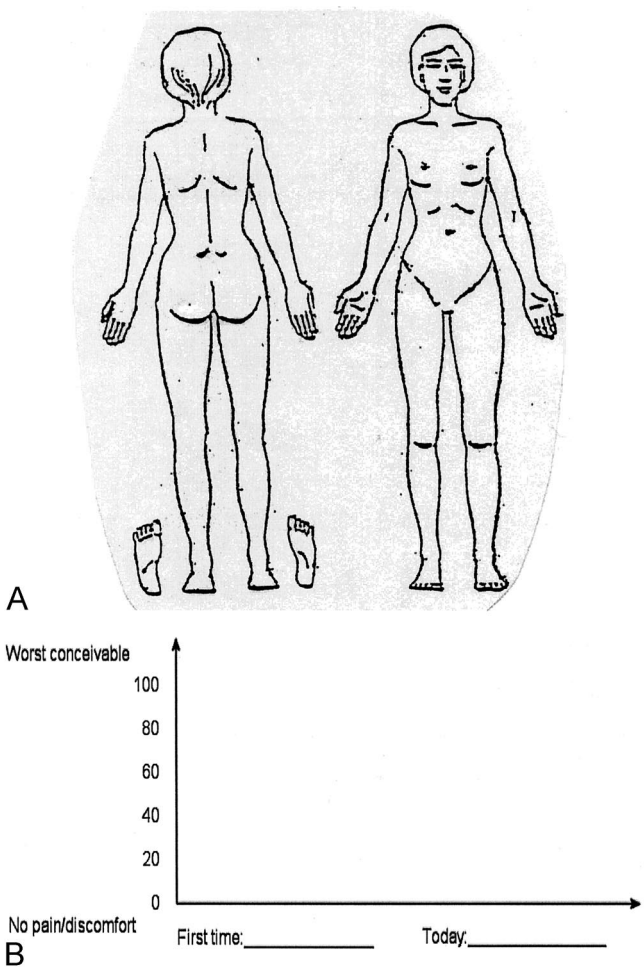

Figure 2. Pain drawing. A. Where and what kind of discomfort? Shadow *all* pain/discomfort this last week, shadow darker where there has been more discomfort. Write by the figure what kind of discomfort: buzzing, tingling, pricking, aching, cramp, *etc.* B. When and how discomforting? How has the pain/discomfort varied since the first time you experienced it?

applicable, immediate treatment. The same information was sent to the family physician in charge. Total time devoted to each patient was about 2 hours.

History. History included a pain drawing, a questionnaire (Appendix I), and an interview. On the pain drawing (Figure 2), the patient noted discomforting areas by shading them with a lead pencil and noted the level of discomfort over time on a chart. The questionnaire contained 22 questions concerning debut, duration, level of pain and dysfunction, and the effect of given treatments. The patient answered the questionnaire with black ink. During the interview, the examiner could add questions as wanted and make comments in red ink to the pain drawing and to the questionnaire.

Physical Examination. The testing technique of the 66 clinical tests, divided into 9 categories, included in the physical examination is presented in Appendix II. Emphasis was placed on neurologic tests because of examiner B's experience that these patients often have neurogenic dysfunction. Both examiners were used to the clinical tests after 6 years of work in the same clinic. However, only the physician performed the sensi-

tivity test in the bimanual way before the study. A pilot study on 24 patients was done to reach unity on how to perform and evaluate history and clinical tests including the bimanual sensitivity test. Two sessions to evaluate performance were held at one-third and two-thirds into the study, respectively.

Statistical Analysis. Interexaminer reliability is expressed as overall agreement in percentages on positive and normal evaluations and by kappa statistics.[7–9] The prevalence of positive findings is presented in percentages. Statistical significance between kappa values, prevalence of positive findings, and agreement on these was accepted if $P < 0.05$ when calculated by the two-sample test for binomial proportions as described by Rosner[10] and Altman.[11] Fisher exact test and Mann-Whitney U test were used to calculate statistically significant differences between the groups concerning the characteristics of the patients.

The kappa value is influenced by the prevalence of positive findings and is attenuated most severely towards low values when the prevalence is either particularly low or high.[12] Therefore, kappa was not calculated when the mean of the examiners prevalence was below 10% or above 90% or when prevalence of one examiner was 0%. Kappa values are classified as follows; <0, worse than chance; 0 to 0.2, poor; 0.21 to 0.4, fair; 0.41 to 0.6, moderate; 0.61 to 0.8, good; and >0.8, very good.[11,13]

■ Results

The patients' characteristics are shown in Table 1. The average patient in both groups was a 43-year-old woman with 4 to 5 years of chronic pain/discomfort, originating from the neck and at a pain level of 60 to 70 on the VAS 100 scale. Neurogenic dysfunction in the area of discomfort was diagnosed in 74% to 92% of the cases at the end of the examiner's assessment. Examiner B noted more neurogenic dysfunction than examiner M (not statistically significant).

There was one statistically significantly difference between the groups. Patients in Group I were diagnosed to have more combined neck/shoulder origin of their discomfort than patients in Group II ($P < 0.05$).

No contraindications to further examination were found in the general status of any patient. In some tests, a decision on whether left, right or both sides showed pathology was considered. These decisions did not alter the kappa value of any category of tests in any significant way, therefore they are not presented.

Interexaminer Reliability

Kappa values altered significantly in 2 of the 66 tests when knowledge of history was added to the physical examination: isometric contractions of the shoulder (0.69/0.27 without/with knowledge, $P < 0.05$) and tenderness of paraspinal muscles (0.04/0.46, $P < 0.01$)(Table 2).

Without knowledge of history, the highest kappa values were noted for neck traction relieves (0.8), shoulder pain on abduction–adduction (0.77), and sensitivity to pain in the neck (0.72). With known history, 7 out of 10 sensitivity tests had values >0.7, as did one tenderness test. The lowest values with and without knowledge of history, indicating less than chance agreement, were found for reflex and hypotrophy tests. Nerve stretch, head movement, and most strength tests showed poor to fair values. The mean kappa value for all tests was 0.42 without and 0.43 with knowledge of history.

Table 1. Characteristics of the Patients

Group	Physical Examination: Without History (I)		Physical Examination: With History (II)	
No. of patients	50		50†	
Age (yrs)				
Mean	42.7		43.5	
Range	18–66		25–66	
Gender (females/males)	30/20		37/13	
Duration of discomfort				
Mean (mos)	57		51	
Range	5 days 25 yrs		9 days–60 yrs	
	No.	%	No.	%
Acute, <1 wk	1	2	0	0
Subacute, 1 wk–3 mos	10	20	18	36
Chronic, >3 mos	39	78	32	64
Pain level (VAS 1–100)				
Median	60		70	
Range	0–100		20–100	
Diagnosis (origin of discomfort) according to the examiner's assessments				
	B/M (no.)	Mean (%)	B/M (no.)	Mean (%)
Neck	29/27	56	39/44	83
Shoulder	7/5	12	3/2	5
Neck and shoulder	13/17	30	7/3	10
Other	1/1	2	1/1	2
	50/50	100	50/50	100
Neurogenic dysfunction in affected area	44/37	81	46/40	86

† Results of one of the examiners physical exam on one patient in Group II were lost, therefore interexaminer data on the clinical tests are computed on 49 instead of 50 patients.
VAS = visual analogue scale.

Prevalence of Positive Findings

With known history, 53 (80%) of the 66 tests showed an increase, 11 (17%) a decrease and 2 (3%) an unchanged prevalence of positive tests (mean of examiners B and M) compared to those tested without knowledge of history (Tables 2 and 3). The increase was statistically significant in 10 tests, 6 of these in sensitivity testing. The highest prevalence of positive findings, 78 to 82% with known history, was noted for palpable tenderness of the spinal processes and paraspinal joints in the lower cervical spine (C4–C7). The sensitivity tests in the corresponding dermatomes (C4–C7) were positive in 45 to 63% and Spurling test in 49 to 68% of the cases when history was known. The tests with the lowest prevalence were the

Table 2. Prevalence of Positive Findings and Interexaminer Reliability Without and With Knowledge of History (n = 49 + 50)

	Prevalence of Positive Findings												Kappa Value				
	Without Examiner			With Examiner			Without With Mean of B + M			Overall Agreement Without With			Without		With		
	B	M		B	M												
Clinical Test	%	%	S	%	%	S	%	%	S	%	%	S	k	SD	k	SD	S
Cervical ROM																	
Ventral flexion	6	12		6	12		9	9		94	90		—	—	—	—	
Extension	22	16		31	16		19	23		82	69		0.42	0.16	0.15	0.15	
Right lateral flexion	18	20		22	20		19	21		74	82		0.16	0.16	0.45	0.15	
Left lateral flexion	24	38		20	38		31	29		70	73		0.31	0.14	0.37	0.14	
Right rotation	18	46†		27	42		32	34		60	63		0.16	0.12	0.2	0.14	
Left rotation	32	56*		31	58*		44	44		68	59		0.39	0.11	0.18	0.14	
Shoulder tests																	
Abduction–adduction	64	70		55	64		67	60		90	82		0.77	0.09	0.62	0.11	
Isometric contraction	32	38		12	34*		35	23		86	73		0.69	0.11	0.27	0.15*	
Tenderness																	
Spinal process C1–C3	42	50		57	60		46	59		80	76		0.6	0.11	0.49	0.13	
Spinal process C4–C7	54	70		78	82		62	80†		72	84		0.42	0.12	0.5	0.15	
Spinal process T1–T3	38	44		55	56		41	56		78	90		0.55	0.12	0.79	0.09	
Spinal process T4–T7	58	54		63	62		56	63		72	76		0.43	0.13	0.47	0.13	
Paraspinal joints C1–C3	58	48		69	62		53	66		66	65		0.32	0.13	0.22	0.14	
Paraspinal joints C4–C7	76	76		84	82		76	83		76	88		0.34	0.15	0.55	0.16	
Paraspinal joints T1–T3	22	52†		39	62*		37	51		70	76		0.41	0.1	0.51	0.12	
Paraspinal joints T4–T7	30	46		43	54		38	48		76	80		0.5	0.12	0.59	0.12	
Neck muscles	22	8		20	16		15	18		82	84		0.32	0.16	0.46	0.16	
Brachial plexus	20	40*		45	46		30	45*		68	61		0.27	0.13	0.22	0.14	
Scapula	30	28		41	30		29	35		66	69		0.17	0.15	0.33	0.14	
Paraspinal muscles	2	18*		18	18		10	18		80	84		−0.04	0.03	0.46	0.16†	
Shoulder	50	66		45	62		58	54		76	89		0.52	0.11	0.38	0.13	
Upper arm	22	12		20	12		17	16		86	84		0.51	0.15	0.4	0.17	
Lateral epicondyle	18	26		27	28		22	27		76	86		0.31	0.15	0.64	0.12	
Medial epicondyle	16	24		18	30		20	24		84	80		0.5	0.15	0.45	0.15	
Lower arm	14	16		12	8		15	10		90	92		0.61	0.16	0.56	0.2	
Thenar	6	6		16	12		6	14		96	88		—	—	0.5	0.18	
Middle hand	10	8		20	18		9	19		94	86		—	—	0.55	0.15	
Hypothenar	4	4		16	14		4	15*		100	90		—	—	0.61	0.15	
Hypotrophy																	
Chin	0	2		2	4		1	3		98	94		—	—	—	—	
Neck	26	2†		14	2		14	8		72	84		−0.04	0.04	—	—	
Neck-shoulder	18	2†		22	2†		10	12		80	80		−0.04	0.03	0.05	0.16	
Shoulder	8	6		14	6		7	10		90	80		—	—	−0.11	0.04	
Upper arm	10	2		0	4		6	2		88	96		—	—	—	—	
Lower arm	4	0		0	0		2	0		96	100		—	—	—	—	
Hand	2	0		8	2		1	5		98	94		—	—	—	—	
Chest	2	12		0	16*		7	8		86	84		—	—	—	—	
Sensitivity to pain																	
Chin	28	28		49	44		28	46*		84	88		0.6	0.13	0.75	0.09	
Neck	36	24		51	38		30	44*		88	84		0.72	0.1	0.71	0.1	
Shoulder	52	46		63	52		49	58		78	73		0.56	0.12	0.46	0.13	
Upper arm	62	50		63	58		56	61		80	88		0.6	0.11	0.78	0.09	
Thumb	32	42		59	52		37	56*		78	90		0.53	0.12	0.79	0.09	
Middle finger	32	36		45	54		34	49*		80	78		0.56	0.12	0.56	0.12	
Little finger	34	32		53	50		33	52*		82	86		0.59	0.12	0.71	0.1	
Axilla	46	44		55	56		45	56		78	86		0.56	0.12	0.71	0.1	
Chest	14	26		31	26		20	28		80	80		0.39	0.15	0.5	0.14	
Foot	26	32		49	46		29	47*		82	86		0.56	0.13	0.71	0.1	
Strength																	
Head flexion	32	24		43	26		28	34		72	67		0.31	0.14	0.28	0.14	
Head lateral flexion	20	26		31	32		23	31		82	78		0.49	0.14	0.48	0.13	
Shoulder elevation	12	32*		20	28		22	24		76	84		0.34	0.13	0.57	0.13	
Arm abduction	16	32		22	30		24	26		76	76		0.36	0.14	0.38	0.14	
Elbow flexion	14	30		22	36		22	29		72	73		0.21	0.14	0.36	0.14	
Elbow extension	28	38		27	30		33	28		66	67		0.24	0.14	0.2	0.15	
Little finger hook	32	46		43	44		39	43		70	59		0.37	0.13	0.18	0.14	

(Table continued)

Table 2. Continued

	Prevalence of Positive Findings												Kappa Value				
	Without Examiner			With Examiner			Without With Mean of B + M			Overall Agreement Without With			Without		With		
	B	M		B	M												
Clinical Test	%	%	S	%	%	S	%	%	S	%	%	S	k	SD	k	SD	S
Reflexes																	
Supraspinatus	0	6		8	12		3	10		94	80		—	—	−0.09	0.04	
Biceps	0	2		4	2		1	3		98	94		—	—	—	—	
Brachioradialis	0	6		8	4		3	6		94	92		—	—	—	—	
Triceps	6	4		6	6		5	6		90	88		—	—	—	—	
Babinski	0	0		0	2		0	1		100	98		—	—	—	—	
Nerve stretch																	
Medianus	10	4		18	8		7	13		94	78*		—	—	0.03	0.15	
Radialis	6	2		16	4		4	10		96	84		—	—	0.11	0.18	
Ulnaris	8	4		16	2*		6	9		92	86		—	—	—	—	
Neck compression/traction																	
Straight compression	26	40		31	44		33	37		70	73		0.34	0.13	0.44	0.13	
Speurling to the right	30	60†		51	68		45	60		66	65		0.37	0.1	0.28	0.14	
Speurling to the left	36	52		49	66		44	58		68	73		0.37	0.12	0.46	0.13	
Traction painful	14	12		27	24		13	25*		90	78		0.56	0.18	0.41	0.15	
Traction relieves	50	56		45	50		53	47		90	82		0.8	0.08	0.63	0.11	
Mean of all	24	28		31	32		26	32		82	81		0.42	0.12	0.43	0.13	

* $P < 0.05$.
† $P < 0.01$.
S = significant difference; k = kappa; SD = standard deviation for kappa; ROM = range of motion.

reflex and nerve-stretch tests with about or less than 10% of positive findings. The mean prevalence of all positive findings for examiner B and M was 31 and 32% respectively.

Bias

A statistically significant difference in prevalence of positive findings between examiners is considered a sign of bias.[14] At interexaminer trial, this was observed in nine tests without and in six tests with knowledge of history (Table 2). There was no statistically significant difference between examiners in the category of sensitivity and reflex tests only.

Mean Kappa Values for Categories of Clinical Tests

There were no statistically significant differences in the mean kappa values of the two examiners for categories of test with and without history. Interexaminer mean kappa values for the 10 sensitivity tests increased from 0.57 to 0.67 with knowledge of history, as did the tenderness tests from 0.4 to 0.49 (Table 3 and Figure 3). The two shoulder tests had a mean kappa of 0.73 without and 0.44 with known history. The other categories of tests were even less affected by history.

■ Discussion

The sensitivity tests we used were the most reliable tests and at the same time the most likely to increase in prevalence with known history. Our results indicate that knowledge of history did not influence reliability of the clinical tests but increased the prevalence of positive findings. Bias in the decision as to what was positive was present in all categories of tests, except the sensitivity and reflex tests. A high degree of positive neurologic findings was observed, especially with sensitivity tests and the neck-compression/traction tests.

Pain diagnostics, *i.e.*, seeking the origin of pain—its anatomic correlation, is of vital importance for the understanding and treatment of any patient.[15] Reliable and valid tests should be sought after in order to secure accurate and effective diagnostic procedures. With knowledge of history, one may expect the reliability of a clinical test to be greater if we believe that there is a tendency to judge clinical findings to fit the history. A single observation of pathology may be interpreted in many ways. Adding other observations to the history or the physical examination usually tends to solidify a working diagnosis as was shown by Vroomen *et al.*[5] However, Ahlbeck[16] showed that the distribution of pain alone (on a pain drawing) was sufficient to predict the level of disc herniation with great accuracy in patients with monoradicular sciatica. Further history or physical examination did not add to the diagnostic accuracy. Also, if history is unreliable and perhaps even influences the interpretation of the clinical findings, then the diagnostic inaccuracy may be increased, as was illustrated by Leclaire *et al.*[4] We have not found any previous study on the possible influence of history on the reliability of a clinical test.

Earlier reliability studies of clinical tests in the examination of the neck/shoulder region have shown moderate to good reliability in identifying the most tender intervertebral joint, assessment of atrophy in the hand, sensitivity for pain and touch, and Spurling test.[17,18]

Table 3. Mean of Positive Findings and Reliability for Categories of Clinical Tests

	Prevalence of Positive Findings						Agreement		Kappa Coefficient			
	Without Examiner			With Examiner			Without %	With %	Without		With	
	B	M	Mean	B	M	Mean			k	SD	k	SD
Clinical Test	%	%	%	%	%	%						
Cervical ROM	20	31	26	23	31	27	75	73	0.29	0.14	0.27	0.15
Shoulder tests	48	54	51	34	49	41	88	78	0.73	0.1	0.44	0.13
Tenderness	30	35	32	39	41	40	79	80	0.4	0.13	0.49	0.14
Hypotrophy	9	3	6	8	5	6	89	89	−0.04	0.04	−0.03	0.1
Sensitivity to pain	36	36	36	52	48	50	81	84	0.57	0.12	0.67	0.11
Strength	22	33	27	30	32	31	73	72	0.33	0.14	0.35	0.14
Reflexes	1	4	2	5	5	5	95	90			−0.09	0.04
Nerve stretch	8	3	6	17	5	11	94	82			0.07	0.17
Neck compression/traction	31	44	38	40	50	45	77	74	0.49	0.12	0.45	0.13

k = kappa; ROM = range of motion; S = significant difference; SD = standard deviation for kappa.

Studies on palpation of tender points, muscle strength, joint play, and range of motion have mainly shown fair to poor reliability.[6,17–25] Standardization of clinical tests[24,26] and a "quick, practical, cost-effective" sensitivity test are being pursued.[27–29]

Our study confirms the reliability ranking of the previously mentioned test categories. Our reliability values may represent an overestimation, compared to the real situation, because we made a pilot study, had two follow-up sessions to coordinate our methods, and intentionally focused on the clinical tests more than what can be expected in a real setting where the examiners probably have less time to evaluate each test. Yet, only our sensitivity test reached values indicating good reliability (with known history). This indicates that most of the clinical tests we used are of questionable value as pertains to reliability. The difference in the prevalence of positive findings between examiners suggests there was bias in the method of evaluating hypotrophy.

A significant difference in the prevalence of positive findings between examiners was found in all categories of tests, except reflex tests (where few tests were positive at all) and sensitivity tests. This indicates that most of the tests we used carry a risk of bias, even the Spurling test, which has proven to be highly specific in detecting radicular pain.[30] Is this a result of the examiners' different educational backgrounds in this study? Previous studies by other health professionals have found similar reliabilities to ours.[6,17–25] Also, the sensitivity test, where we had no common experience before the study due to our different methods of performing the test, turned out to be the most reliable and least differing in prevalence of positive findings. This suggests that education was not the determining factor. Rather, we believe that the reliability of a clinical test depends more on how difficult the test is to perform in a standardized manner.

The history questionnaire we used was not standardized nor validated and included open-ended questions. This limits the possibility to compare our results with other similar studies. Yet, in a clinical setting, open questions are the rule, and we tried to assimilate the real setting. The observation that the prevalence of positive findings increased significantly with known history in 6 out of 10 of our sensitivity tests but in only 4 of all other

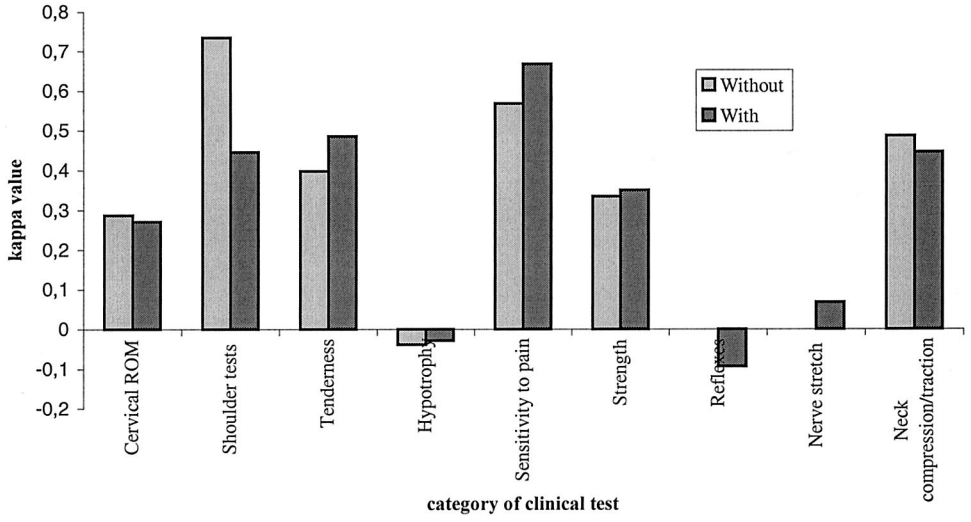

Figure 3. Mean kappa values for categories of clinical tests without and with knowledge of history.

tests is interesting. Is our sensitivity test not specific or insensitive, or is there something in the history that makes us perform this test differently when we know what to look for? We believe the latter to be true as we found not only the mean prevalence of all sensitivity tests increased with knowledge of history, from 36% to 50%, but also their mean kappa value, from 0.57 to 0.67 (Table 3). Our hypothesis is that patients overlooked slight differences in sensitivity but that knowledge of history made us question their response and ask again if there was a sensitivity difference in areas where we could expect it. Actually, our bimanual sensitivity test carried the least risk for bias and was the most reliable (highest mean of kappa with known history). This may be an answer to the sought-after "quick, practical, cost-effective" test[27–29] that can be performed by all health workers.

The fact that about half of all the neck-compression/traction and sensitivity tests were positive with a known history, whereas pathology in strength tests was observed in about one-third and one-tenth or less in nerve stretch, reflex, and hypotrophy tests (Table 3) leads to the question: which level of neurogenic dysfunction and pain is the most valid? We believe that the levels of pathology found in sensitivity and neck-compression/traction tests may be the most valid because their kappa values were considerably higher (moderate to good), whereas the strength tests reached fair and the other neurologic tests showed poor values. Also, Viikari-Juntura *et al*, in their studies of neck compression and traction tests, found these tests highly specific but with a sensitivity of about 30%.[30] We therefore believe that the level of observed neurogenic dysfunction in our sensitivity and neck-compression and traction tests represents a minimum of what is true in our population.

Was our study population a select segment with cervical radiculopathy unrepresentative of family practitioners patients? We believe it was representative. The primary care units were repeatedly reminded to refer *all* consecutive patients seeking help for neck and/or shoulder problems according to the preset criteria. Our impression is that they did so. If this is true, it prompts the question: are many so-called tense and stiff muscles, trigger points, and shoulder impingement problems in reality neurogenic dysfunction syndromes originating from the spine? This idea is supported by our observation that patients with known history had sensitivity signs of neurogenic dysfunction in about 60% of the cases in the C4–C7 dermatome areas—the same areas where they experienced problems—and palpable tenderness in the spine in the region where these nerves originate (C4–C7) in about 80% of the cases. Further studies are needed to confirm the reliability and prevalence of our findings and to assess the validity (sensitivity and specificity) of these clinical tests.

■ Conclusions

This study of primary health care patients with neck and/or shoulder problems has evaluated the reliability of clinical tests, the prevalence of positive findings, and the impact of history on these parameters. Our simple, bimanual, sensitivity test was the only test that showed good reliability and no bias. This test should now undergo further reliability and validation studies. Head movement and strength tests showed poor or fair reliability and/or a risk of bias, signifying a lower diagnostic value until standardization methods prove otherwise. History had little or no impact on the reliability of the tests. However, the prevalence of positive findings increased with knowledge of history, perhaps because of the increased awareness of where to look for it.

Finally, neurogenic dysfunction, in one or more of the affected areas, was observed in four-fifths of the patients and may be a greatly underestimated reason for problems in the neck/shoulder region. Therefore, we recommend that it be tested for with careful, bimanual, sensitivity testing techniques.

■ Key Points

- Only a bimanual method of sensitivity testing reached good kappa values and was exempt from bias.
- The reliability of clinical tests used in our study was the same with and without knowledge of history, but the prevalence of positive findings increased with knowledge of history.
- A substantial number of positive findings in neurologic tests suggests that neurogenic dysfunction is common in primary care patients with neck/shoulder problems.

■ Appendix I: Questionnaire

Patient name _____
Social security # _____
Address _____
Telephone/fax # _____
Examination date _____
Birth date _____
Sex (male/female) _____
Social status _____
Occupation _____
On sick leave since _____
Tobacco use (yes/no) _____
Referral from (name of family physician) _____

History of neck/shoulder discomfort (write and/or circle the answer)

Why do you seek help?

_____

When did the discomfort you seek help for start?

_____

3a. Was it an accident? yes no
b. If yes, what happened?

_____

4. How did the discomfort start?

_____

5. Have you had similar discomfort before? yes no
6. In what body part(s) do you experience discomfort?
   a. head right left
   b. neck right left
   c. shoulder right left
   d. arm right left
   e. hand/finger(s) right left
   f. Discomfort in some other part(s) of the body?

_____

7. Do you experience constant discomfort day as night? yes no
8. What increases your discomfort?
   a. sudden effort like coughing or laughing yes no
   b. turning your head yes no
   c. shoulder movement yes no
   d. other

_____

9. Do you experience relief from your discomfort by laying down? yes no
10. What decreases your discomfort?

_____

11. Is your discomfort associated with?
   a. wryneck yes no
   b. headache yes no
   c. dizziness yes no
   d. tingling or numbness in lower arm yes no
   e. tingling or numbness in hand/finger yes no

12. Draw a stroke on the following lines to illustrate the last week's experience of:
   a. Pain
      no = 0_____ 10 = worst conceivable
   b. Problem to sleep
      no = 0_____ 10 = worst conceivable
   c. Problem to work. With what?

      _____
      no = 0_____ 10 = worst conceivable
   d. Problem at leisure time. With what?
      no = 0_____ 10 = worst conceivable

13. Quality of life? (assess by drawing a stroke on the following line)
      no = 0_____ 10 = best conceivable
14a. Assessed by (*e.g.*, family physician, orthopaedist):

_____

   b. With (*e.g.*, blood test, radiology, EMG):

_____

   c. Have you received a trustworthy explanation to your discomfort? yes no
   d. Which explanation?

_____

15a. Have you been x-rayed, if so what and when?

_____

   b. Do you bring x-ray replies? yes no
16a. Treated by (*e.g.*, physician, physiotherapist, alternative medicine): _____
   b. With what? (*e.g.*, ultrasound, electric current, hot bath): _____

17. Previous and other medical discomfort *e.g.*, operation/hospital stay/current disease:

_____

_____

18. Current medication:

_____

19. What is your greatest discomfort?

_____

20. What do you believe is the reason for your discomfort?

_____

Diagnosis: _____code 1 2 3 4 5 with/without nerve dysfunction

■ Appendix II: Technique of the Physical Examination

A general status was first considered while the patient undressed the upper body. Torticollis and, if so, the position of the head were noted. Problems in moving, antalgic positioning, mental or speech disorder, skin disease or other notable, and physical or mental deficiency were noted. Patients were then examined in sitting position with undressed upper body and hands resting on their legs. All tests presented were measured subjectively, including angles where we had tested our measuring ability with a set square on some pilot patients. Pathology, yes or no, was first considered and then, if pathology, right and/or left side was noted. Uncertain pathology was considered as no pathology. In some tests, the positive findings were graded. In this paper, graded findings are not presented. The 66 clinical tests were divided into the following 9 categories.

Cervical range of motion. Active, extension, ventral and lateral flexion, and rotation to each side of the head until pain or stiffness stopped the movement was observed while standing in front of the patient. Normal movement was defined in accordance with the limits suggested by Viikari-Juntura *et al*[30]: ventral flexion and extension 30°, lateral flexion to each side 20°, and rotation to each side 60° or more. Movement inhibited before reaching the limits was considered positive, and the area of any pain was asked for and noted.

Shoulder tests. Active abduction to 180° and adduction to 30° with thumb pointing upwards were tested subjectively, standing in front of the patient. Limited motion and/or distinct pain within the given range were considered positive, and the area of pain was noted. Isometric contraction of shoulder muscles was tried with the examiner standing behind the patient with hands on the patients arms held in 90° flexion in the elbow and thumb pointing upwards. While giving resistance to both arms, the examiner would ask the patient to exert force bilaterally in one of eight directions at the time: lower arm up, down, external rotation, internal rotation, upper-arm abduction, flexion, extension, and adduction. Distinct pain was considered positive and was specified as to side and location.

Tenderness. With the examiner standing behind the patient, a mild to moderate pressure with one or two

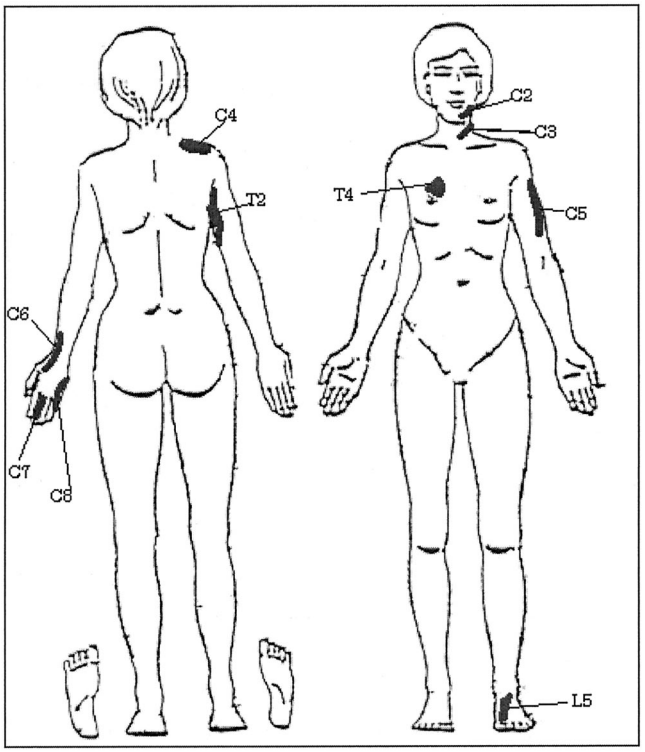

Figure 4. Indicator areas for dermatomes. The delimitation between the dermatomes differs between different atlases especially in the proximal areas (C2–C5).

fingers was exerted on each one of the specified structures in Table 2, bilateral where feasible, and asking the patient for pain response. Distinct pain was considered positive and noted.

Neurologic Tests

Hypotrophy was noted in the areas described in Table 2 representing myotomes; chin C2, neck C3, neck/shoulder C3–C4, shoulder C4–C6, upper arm C5–C6, lower arm C6–C7, hand C6–C8, and chest C6-T4.

Sensitivity to pain was tested by using two pinwheels drawn slowly, with no pressure other than their own weight, bilaterally, simultaneously, over indicator areas for dermatomes noted in Table 2 and shown in Figure 4. The patient was asked if he/she experienced a difference from side to side or from chin to foot. Increased or decreased sensitivity in an area was noted.

Strength deficiency was tested by the examiner standing behind the patient and asking him or her to resist force from the examiners hand in a certain direction representing action in different myotomes; head flexion C2–C3, head lateral flexion C3–C4, shoulder elevation C4–C5, arm abduction C5–C6, elbow flexion C5–C6, elbow extension C6–C7, and little finger hook C7–C8.

Except for head movements, all tests were done simultaneously on both sides. Force was applied for about 5 seconds per test. Decreased strength in one or both sides was noted.

Reflexes were tested on one side at a time with a reflex hammer. Asymmetry was considered positive as well as a weak or strong reaction.

Nerve stretch was performed for the median, radial and ulnar nerve, using one arm at a time and noting pain response. Pain response in the arm and brachial plexus was noted.

Neck compression/traction. Compression and traction of the neck was done with the examiner standing behind the patient with hands on top of the patients head and exerting increased pressure, with the head in different positions, respectively lifting the head with hands underneath each maxilla and with thumbs on the back of the head as suggested by Viikari-Juntura *et al*.[30] If the patient expressed pain, further compression/traction was immediately stopped and the pain was specified as to side and area. Relief at the traction of the head was specified as to side and area of relief.

References

1. Koran LM. The reliability of clinical methods, data and judgments (first of two parts). *N Engl J Med* 1975;293:642–6.
2. Bjerregaard B, Brynitz S, Holst-Christensen J, et al. The reliability of medical history and physical examination in patients with acute abdominal pain. *Methods Inf Med* 1983;22:15–8.
3. Michel A, Kohlmann T, Raspe H. The association between clinical findings on physical examination and self-reported severity in back pain. Results of a population-based study. *Spine* 1997;22:296–303; discussion 304.
4. Leclaire R, Esdaile JM, Jequier JC, et al. Diagnostic accuracy of technologies used in low back pain assessment. Thermography, triaxial dynamometry, spinoscopy, and clinical examination. *Spine* 1996;21:1325–30; discussion 31.
5. Vroomen PC, de Krom MC, Knottnerus JA. Consistency of history taking and physical examination in patients with suspected lumbar nerve root involvement. *Spine* 2000;25:91–6; discussion 97.
6. Sobel JB, Sollenberger P, Robinson R, et al. Cervical nonorganic signs: a new clinical tool to assess abnormal illness behavior in neck pain patients: a pilot study. *Arch Phys Med Rehabil* 2000;81:170–5.
7. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *Br Med J* 1992;304:1491–4.
8. Haas M. Statistical methodology for reliability studies [see comments]. *J Manipulative Physiol Ther* 1991;14:119–32.
9. Haas M. How to evaluate intraexaminer reliability using an interexaminer reliability study design. *J Manipulative Physiol Ther* 1995;18:10–5.
10. Rosner B. *Fundamentals of Biostatistics*. 5th ed. Doxbury; 2000.
11. Altman DG. *Practical Statistics for Medical Research*. Chapman and Hall; 1999.
12. Walter SD. Measuring the reliability of clinical data: the case for using three observers. *Rev Epidemiol Sante Publique* 1984;32:206–11.
13. Strender LE, Sjoblom A, Sundell K, et al. Interexaminer reliability in physical examination of patients with low back pain. *Spine* 1997;22:814–20.
14. Cambell MJ, Machin D. *Medical Statistics: A Common Sense Approach*. 2nd ed. Wiley; 1993.
15. Wilson L, Hall H, McIntosh G, et al. Intertester reliability of a low back pain classification system. *Spine* 1999;24:248–54.
16. Albeck MJ. A critical assessment of clinical diagnosis of disc herniation in patients with monoradicular sciatica. *Acta Neurochir* 1996;138:40–4.
17. Viikari-Juntura E. Interexaminer reliability of observations in physical examinations of the neck. *Phys Ther* 1987;67:1526–32.
18. Hubka MJ, Phelan SP. Interexaminer reliability of palpation for cervical spine tenderness. *J Manipulative Physiol Ther* 1994;17:591–5.
19. Deboer KF, Harmon R Jr, Tuttle CD, et al. Reliability study of detection of somatic dysfunctions in the cervical spine. *J Manipulative Physiol Ther* 1985;8:9–16.
20. Fjellner A:Bexander C, Faleij R, et al. Interexaminer reliability in physical examination of the cervical spine. *J Manipulative Physiol Ther* 1999;22:511–6.
21. Frese E, Brown M, Norton BJ. Clinical reliability of manual muscle testing. Middle trapezius and gluteus medius muscles. *Phys Ther* 1987;67:1072–6.
22. Nansel DD, Peneff AL, Jansen RD, et al. Interexaminer concordance in

detecting joint-play asymmetries in the cervical spines of otherwise asymptomatic subjects [see comments]. *J Manipulative Physiol Ther* 1989;12:428–33.
23. Smedmark V, Wallin M, Arvidsson I. Inter-examiner reliability in assessing passive intervertebral motion of the cervical spine. *Man Ther* 2000;5:97–101.
24. Strender LE, Lundin M, Nell K. Interexaminer reliability in physical examination of the neck. *J Manipulative Physiol Ther* 1997;20:516–20.
25. Van Suijlekom HA, De Vet HC, Van Den Berg SG, et al. Interobserver reliability in physical examination of the cervical spine in patients with headache. *Headache* 2000;40:581–6.
26. Breen A. The reliability of palpation and other diagnostic methods. *J Manipulative Physiol Ther* 1992;15:54–6.
27. Bell-Krotoski J. Advances in sensibility evaluation. *Hand Clin* 1991;7:527–46.
28. Bell-Krotoski J, Weinstein S, Weinstein C. Testing sensibility, including touch-pressure, two-point discrimination, point localization, and vibration. *J Hand Ther* 1993;6:114–23.
29. Bell-Krotoski JA, Buford WL Jr. The force/time relationship of clinically used sensory testing instruments. *J Hand Ther* 1997;10:297–309.
30. Viikari-Juntura E, Porras M, Laasonen EM. Validity of clinical tests in the diagnosis of root compression in cervical disc disease. *Spine* 1989;14:253–7.