SPINE Volume 29, Number 19, pp E413–E425

Reliability of Spinal Palpation for Diagnosis of Back and Neck Pain

A Systematic Review of the Literature

Michael A. SeffÞnger, DO,* Wadie I. Najm, MD, Shiraz I. Mishra, MD, PhD,à Alan Adams, DC, MS,¤ Vivian M. Dickerson, MD,¤ Linda S. Murphy, MLIS,¦ and Sibylle Reinsch, PhD**

Study Design. A systematic review.

Objectives. To determine the quality of the research and assess the interexaminer and intraexaminer reliability of spinal palpatory diagnostic procedures.

Summary of Background Data. Conßicting data have been reported over the past 35 years regarding the reliability of spinal palpatory tests.

Methods. The authors used 13 electronic databases and manually searched the literature from January 1, 1966 to October 1, 2001. Forty-nine (6%) of 797 primary research articles met the inclusion criteria. Two blinded, independent reviewers scored each article. Consensus or a content expert reconciled discrepancies.

Results. The quality scores ranged from 25 to 79/100. Subject description, study design, and presentation of results were the weakest areas. The 12 highest quality articles found pain provocation, motion, and landmark location tests to have acceptable reliability (K 0.40 or greater), but they were not always reproducible by other examiners under similar conditions. In those that used kappa statistics, a higher percentage of the pain provocation studies (64%) demonstrated acceptable reliability, followed by motion studies (58%), landmark (33%), and soft tissue studies (0%). Regional range of motion is more reliable than segmental range of motion, and intraexaminer reliability is better than interexaminer reliability. Overall, examinersÕ disciplinerall,a

regional and segmental motion, paraspinal soft tissue abnormalities, and tenderness on provocation. The ability to arrive at an accurate palpatory assessment depends mainly on the validity and reliability of the palpatory tests used.

Although validity and reliability are often used interchangeably in the literature, they are not synonymous. Validity is the accuracy of a measurement of the true state of a phenomenon.[9] Reliability measures the concordance, consistency, or repeatability of outcomes.[10]

Over the past 30 years, scientists with diverse professional training have investigated the validity and/or reliability of spinal palpatory tests used to diagnose nonspecific back pain.[11–13] Several narrative reviews of the literature on spinal palpatory diagnostic procedures have been published.[14–17] However, only two systematic reviews of reliability studies of spinal palpatory tests have been published. One is a limited review of chiropractic literature on palpatory diagnostic procedures for the lumbar-pelvic spine[18]; the other[19] focused on the reliability of sacroiliac joint palpatory tests. The reliability of spinal palpatory diagnostic procedures for neck and back problems remains unclear. There is no comprehensive systematic review of the literature on the reliability of cervical, thoracic, and lumbar spinal palpatory diagnostic procedures.

The authors performed a systematic review of original research articles, from all disciplines, published in peer-reviewed journals in order to assess the quality of the literature and answer the clinical question:" What is the intra- and interexaminer reliability of spinal palpatory diagnostic procedures?

■ Materials and Methods

A multidisciplinary team conducted the systematic review at the Susan Samueli Center for Complementary and Alternative Medicine (University of California, Irvine), between October 2001 and December 2002. The research team included expertise in database searches, clinical research, evidence-based medicine, research design, and statistics methodology. The clinicians represented content area experts in osteopathic, chiropractic, and family medicine/primary care.

A comprehensive strategy, including the exploration of 13 online databases and a manual search of appropriate literature, guided the search for pertinent articles that addressed the study question. Articles were limited to human studies published in peer-reviewed journals or dissertations published between January 1, 1966 and October 1, 2001. All databases were searched using a basic search template. When appropriate, minor modifications to the basic search template were made to optimize the search strategy in individual databases. The 13 databases included: PubMed MEDLINE, MANTIS, MD Consult, Web of Science, EMBASE, CINAHL, BIOSIS Preview, Index to Chiropractic Literature, OSTMED, OCLC FirstSearch, Digital Dissertation, PEDro, and Cochrane. Selection of these databases was determined by the availability of online resources accessible from our institution and affiliated institution libraries, as well as potential inclusion of articles from osteopathic medicine, allopathic medicine, chiropractic medicine, manual medicine, and physical therapy. The manual search included gleaning references cited in studies selected from the online search, and

reliability studies. Thus, the total number of studies included in the 49 articles is 53. Descriptions of the characteristics of the studies are summarized in Table 1.

Paired reviewers initially disagreed on the quality score of 16 (33%) of the 49 articles. Quality scores of the 49 articles ranged from 25 to 79 of 100. The authors compared quality scores of articles in the top quartile (67.5–79) to those in the bottom quartile (25–47). No correlation between quality score and year of publication, examiners' disciplines (clinical degree or specialty training), or procedure evaluated was found. All studies were lacking in description of subjects. Study design, description of study conditions and examiner's professional training, data analysis, and presentation of results were the weakest areas in the lower quality studies.

Interestingly, symptomatic (back or neck pain) subjects were recruited only in 14 (26%) of the 53 studies, and both symptomatic and asymptomatic subjects were recruited in only 9 of 53 (17%). Additionally, two studies assessed the effect of hypertensive subjects on the reliability of palpatory findings.[41,42]

The authors synthesized the data only from the higher quality articles (quality score 67.5 of 100 or greater). Most (two thirds) of the higher quality articles used the more rigorous kappa or weighted kappa measure of as-

Table 2. Quality Scores, Study Characteristics, and Intraexaminer and Interexaminer Reliability for Motion Palpation Tests

Study	Quality Score	Examiners, Subjects	Type of Reliability, Spinal Motion Tests, and Results	Interpretation*
Strenderet al[48]	79.0	2 PT; 25 Sx, 25 ASx subjects	InterEx, cervical segmental K 0.09–0.15; 26–44% agreement	Low reliability
Schopset al[49]	77.5	5 Physicians; 20 Sx subjects	InterEx, cervical and thoracic segmental K 0.6–0.8 for 1st 2 examiners; 0.02–0.24 for all 5	Low to high reliability, examiner dependent
Fjellneret al[44]	74.0	2 PT; 47 (11 Sx and 35 ASx, 1 UMS) subjects	InterEx, cervical and thoracic, regional and segmental Regional ROM: Kw 0.4 in 6 of 8 tests except for rotation; Regional end-feel motion tests: Kw 0.4 in 3 of 8 tests Passive segmental tests: Kw 0.4 in 5 of 58 exams	Regional ROM, except for rotation, some end-feel and some segmental motion tests: medium reliability; most end-feel and segmental exams had low reliability
Loveet al[45]	72.0	8 DC students; 32 ASx subjects	IntraEx and InterEx, thoracic and lumbar segmental IntraEx: Pearson r 0.302–0.6856 InterEx: Index of Association statistic (R)0.023–0.0852	IntraEx more reliable than InterEx
Johnstonet al[42]	71.0	3 DO; 307 (153 hypertensive) subjects	InterEx, cervical and thoracic segmental Higher level of InterEx agreement in subsample with more hypertensives (χ^2 27.75, df 1, P 0.001)	More reliable in hypertensive subjects
Lundbergt al[52]	68.0	2 PT; 150 UMS subjects	InterEx, thoracic and lumbar segmental K (w) 0.42–0.75	Medium to high reliability
Keatinget al[46]	67.5	3 DC; 46 (21 Sx and 25 ASx) subjects	InterEx, thoracic and lumbar segmental Active motion palpation mean K 0.00–0.25; Passive motion palpation mean K 0.03–0.23	Low reliability; no significant differences between Sx and ASx subjects
Johnstonet al[41]	67.0	3 DO (2 students); 132 Asx (some hypertensive) subjects	InterEx, cervical and thoracic segmental 39.5% observed vs. 26.0% expected agreement, P 0.05	More reliable in hypertensive subjects
Maher et al[66]	66.0	6 PT; 90 Sx subjects	InterEx, lumbar segmental 13–43% agreement ICC 0.4 0.73	Low reliability
Grantet al[67]	65.5	4 DC students; 60 UMS subjects	IntraEx and InterEx, lumbar segmental IntraEx: 85–90% agreement InterEx: 66.7% agreement	IntraEx more reliable than InterEx
Haaset al[68]	64.5	2 DC; 73 (48 Sx and 25 ASx) subjects	IntraEx and InterEx, thoracic segmental IntraEx: K 0.43–0.55 InterEx: K 0.14 (segmental level) and K 0.19 (segmental restriction)	IntraEx: medium reliability; InterEx: low reliability; no difference between Sx and ASx subjects
Deboeret al[69]	64.5	3 DC; 40 Asx subjects	IntraEx and InterEx, cervical segmental IntraEx: 45–75% agreement; K (w) 0.01–0.76 InterEx: 21–58% agreement; K (w) 0.03–0.45	IntraEx: low reliability, except one value was high at C1–C2; InterEx: low to medium reliability, more reliable at C6–C7 than C1–C5
Phillipset al[70]	63.0	2 PT; 72 (63 Sx and 9 ASx) subjects	InterEx, lumbar segmental 55–100% agreement K (w) 0.15–0.32	Low reliability; includes quality of motion and end-feel or tissue response during motion testing
Strenderet al[53]	62.5	2 PT; 50 Sx subjects	InterEx, lumbar regional and segmental Regional ROM: 87–94% agreement; K 0.43–0.74 Segmental: 72–88% agreement; K 0.38–0.75	Regional ROM—extension and lateral bend: medium reliability Segmental: medium to high reliability at lumbosacral joint andone segment above it
Strenderet al[53]	62.5	2 MD; 21 Sx subjects	InterEx, lumbar regional and segmental Regional ROM: 88–96% agreement; K 0.11–0.35 Segmental: 48–86% agreement; K 0.08–0.24	Regional ROM—extension and lateral bend: low reliability Segmental: low reliability
Mastrianiet al[71]	61.5	3 PT; 16 Sx subjects	InterEx, lumbar segmental L3–L4: 70–73% agreement; All segments combined: 66–76% agreement	Low reliability; more reliable at L3–4
Bolineet al[72]	60.0	2 DC (1 student); 50 (23 Sx and 27 ASx) subjects	InterEx, lumbar segmental K 0.05–0.31	Low reliability; no significant differences between Sx and ASx subjects
Inscoeet al[73]	59.0	2 PT; 6 Sx subjectsbend:		

The majority of spinal palpatory diagnostic tests demonstrated low reliability. Data from the higher quality studies (quality score 67.5 of 100 or greater) showed acceptable reliability for the following spinal palpatory diagnostic procedures: 1) interexaminer regional range of motion of the cervical spine[44]; 2) intraexaminer thoracic and lumbar segmental vertebral motion tests[45]; 3) interexaminer pain provocation at a) L4–L5 and L5–S1,[46] b) lumbar paraspinal myofascial trigger points (between trained examiners only);[47] c) the cervical spine;[48,49]

Table 3. Quality Scores, Study Characteristics, Intra- and InterEx Reliability for Pain Provocation Tests

Study	Quality Score	Examiners, Subjects	Type of Reliability, Spinal Region, Pain Provocation Test, and Results	Interpretation*
Strenderet al[48]	79.0	2 PT; 50 (25 Sx and 25 ASx) subjects	InterEx, cervical digital pressure K 0.31–0.52; 58–76% agreement	Low to medium reliability; no difference between Sx and ASx subjects
Schopset al[49]	77.5	5 Physicians; 20 Sx subjects	InterEx, cervical and thoracic digital pressure K 0.2–0.6 C-spine; K 0.6–0.75 T1; K 0.2–0.75 muscles	C: low to medium reliability T1: medium reliability Muscles: low reliability, except SCM which had medium reliability
Hsiehet al[47]	69.0	8 examiners: 1 expert MD; 4 trained: 2 DC, 1 DO and 1 MD; 4 untrained: 2 DC and 2 MD; 52 (26 Sx and 26 ASx) subjects	InterEx, lumbar referred pain upon digital pressure on trigger point InterEx: Trained K 0.435; Untrained K 0.320 Agreement with expert: Trained K 0.337; Untrained K 0.292	Low reliability overall except for medium reliability between trained examiners, but not with expert
Lundbergt al[52]	68.0	2 PT; 150 UMS subjects	InterEx, thoracic and lumbar digital pressure L4–L5: K 0.71 L5-S1: K 0.67	L4–L5 and L5-S1: medium reliability Data for thoracic and other lumbar segments not reported
Keatinget al[46]	67.5	3 DC; 46 (21 Sx and 25 ASx) subjects	InterEx, thoracic and lumbar bony and soft tissue digital pressure K 0.22–0.42 for soft tissue pain; K 0.34–0.65 for osseous pain (mean 0.48)	Low to medium reliability; L4–5 and L5-S1 had greater concordance for osseous pain (mean K 0.6); no signi cant difference between Sx vs. ASx subjects
Maher et al[66]	66.0	6 PT; 90 Sx subjects	InterEx, lumbar predictive reliability of subject pain upon palpation 27–57% agreement; ICC: 0.27–0.85	Low to occasionally reliable
McPartlandet al[88]	66.0	2 DO; 18 (7 Sx and 11 ASx) subjects	InterEx, cervical digital pressure on Strain-counterstrain tenderpoints Sx subjects: 72.7% agreement; K 0.45; ASx subjects: 59.43% agreement; K 0.19	Medium reliability in Sx subjects; low reliability in ASx subjects
McPartlandet al[88]	66.0	18 DO students; 18 ASx subjects	InterEx, cervical digital pressure on Strain-counterstrain tenderpoints 64.2% agreement; K 0.2	Low reliability
Deboeret al[69]	64.5	3 DC; 40 ASx subjects	IntraEx and InterEx, cervical digital pressure IntraEx: C1–C3: 55–80% agreement, Kw 0.3–0.56; C4–C7: 60–68% agreement, Kw 0.2–0.43; InterEx: C1–C3: 43–66% agreement, Kw 0.08–0.48; C4–C7: 34–53% agreement, Kw - 0.04–0.18	Both IntraEx and InterEx: low to medium reliability; IntraEx more reliable than InterEx reliability; both more reliable at C1–C3 than C4–C7
Strenderet al[53]	62.5	2 PT; 50 Sx subjects	InterEx, lumbar paravertebral digital pressure and segmental, lateral bend, extension–exion, foramen compression passive motion tests 78–98% agreement; K 0.27 for paravertebral tenderness; K 0.43–0.76 for regional lateral bend, exion, extension pain and segmental lumbosacral and one segment above lumbosacral pain; foramen compression test: 94% agreement Sensibility at L4: 98% and L5: 97% agreement; all 3 tests: prevalence 10%†	Training made no difference; paravertebral tenderness: low reliability; segmental, lateral bend, extension and exion pain, foramen compression test, and sensibility at L4 and L5 upon digital pressure all had medium to high reliability
Strenderet al[53]	62.5	2 MD; 21 Sx subjects	InterEx, lumbar paravertebral digital pressure, and segmental, lateral bend, extension and exion, foramen compression passive motion tests Lateral bend pain: 73% agreement; K 0.06. Extension and exion pain: 86% agreement; K 0.71. Paravertebral tenderness: 76%, K 0.22. Lumbosacral segment and one above it tenderness: 71% agreement; K 0.40 Foramen compression test: 98% agreement; sensibility at L4 and L5: 100% agreement; prevalence 10%†	Lateral bend pain and paravertebral tenderness: low reliability Extension and exion pain: medium reliability Lumbosacral segment and one segment above it medium reliability Foramen compression test and sensibility at L4–L5: high reliability
Hubkaet al[89]	62.0	2 DC; 30 Sx subjects	InterEx, cervical digital pressure 76.6% agreement; K 0.68	Medium reliability
Boline et al[72]	60.0	2 DC (1 student); 50 (23 Sx and 27 ASx) subjects	InterEx, lumbar digital pressure Sx subjects: L2–L3 and L3–L4 only: 96% agreement; K 0.65; Other lumbar levels: 81% (L5–S1)-91% (T12–L1 and L1–L2) agreement; K 0–0.06 Both ASx and Sx subjects combined: 96% agreement; K 0.03–0.37 at T12–L2 and L3–S1; K 0.49 at L2–L3	Sx subjects at L2–L3 and L3–L4: medium reliability; rest of L-spine: low reliability With both Sx and Asx subjects at L2–L3: medium reliability; rest of L-spine: low reliability
Viikari-Junturaet al[90]	58.5	1 MD and 1 PT; 52 Sx subjects	InterEx, cervical (C5–C8) digital pressure tenderness, sensitivity and foramen compression passive motion test K 0.24–0.56 for tenderness to palpation; K 0.41–0.64 for sensitivity testing; K 0.28–0.77 for segmental foramen compression test for radiculopathy	Tenderness: low to medium reliability; sensitivity: medium reliability Foramen compression test: low to high reliability; most reliable for radicular symptoms to the forearm
Nice et al[91]	52.0	12 PT; 50 Sx subjects	InterEx, lumbar trigger point digital pressure 76–79% agreement, K 0.29–0.38	Low reliability; improved reliability noted when examiners followed proper technique per protocol and subjects reported Sx immediately prior to examination
Boline et al[92]	43.0	3 DC; 28 Sx subjects	InterEx, lumbar osseous and soft tissue digital pressure Osseous pain provocation: 79–96% agreement, K 0.48–0.90; Soft-tissue pain provocation: 75	

There are informative trends noticeable among the higher quality quartile studies that used the same statistical analysis. In those studies that used kappa statistics, a higher percentage of the pain provocation studies (7 of 11; 64%) demonstrated acceptable reliability followed by motion studies (7 of 12; 58%), landmark studies (1 of 3; 33%), and soft tissue studies (0 of 11; 0%). No spinal

Table 5. Quality Scores, Study Characteristics, Intra- and InterEx Reliability for Landmark Tests

Study	Quality Score	Examiners Subjects	Type of Reliability, Spinal Region, Landmark Test, and Results	Interpretation*
Downeyet al[50]	72.0	6 PT; 60 Sx subjects	InterEx, lumbar location of nominated lumbar spinal level K 0.44–0.88 for agreement on one nominated level; Kw 0.86–0.98 (scale and criteria not reported)	Medium to high reliability; selected examiners trained and educated in manipulative therapy, and accepted a range of determinations as being concordant; improved agreement by

cedures. In a systematic review of the content validity of spinal palpatory tests, the authors found that pain scales were one of only a few validated instruments that can be used in these types of studies.[56]

The spinal examination, with its small joints and limited mobility, may be more dif cult for most clinicians than more prominent joints. The larger joints of the extremities fare slightly better (i.e., physical therapists assessing shoulder motion restriction, kappa 0.62–0.76).[57] However, the smaller joints of the extremities, like the vertebral spine, are less reliable (i.e., kappa 0.24–0.60 among rheumatologists palpating for hard tissue enlargement of hand and nger joints).[58]

Evaluation of the reliability of physical examination procedures in general poses a number of methodologic challenges. Examiner bias and inconsistency create variability in procedures. Although palpation for pedal pulses has medium to high reliability (kappa 0.54–0.87),[59] many physical examination procedures used commonly in clinical practice have low to medium reliability.[60,61] This includes lung auscultation (kappa 0.32 for bronchial breath sounds and 0.51 for wheezes)[62] and heart auscultation (31%–72% agreement among physicians).[63]

The primary research articles on the reliability of spinal palpatory procedures are dif cult to compare due to variability in the palpatory tests, terminology, research design, study conditions, and statistical analysis used. The

Table 7. Reliability Articles Weighted Mean Quality Scores

Reliability Article listed by author(s) (year of publication)	Subjects (18)*	Examiners (25)*	Condition (25)*	Analysis (25)*	Results (7)*	Overall (total 100)*
Strenderet al (1997)[48]	5.0	25.0	25.0	17.0	7.0	79.0
Schopset al (2000)[49]	5.5	25.0	23.5	18.0	5.5	77.5
Fjellner (1999)[44]	5.0	17.0	21.0	25.0	6.0	74.0
Rouwmaatet al (1998)[34]	4.0	17.0	20.5	25.0	7.0	73.5
Downeyet al (1999)[50]	3.0	17.0	21.0	25.0	6.0	72.0
Loveet al (1987)[45] (1987)	4.0	25.0	21.0	18.0	4.0	72.0

quality scoring instrument helped to evaluate the relative value of their results. The quality assessment form can also provide a template with which future higher quality reliability studies can be designed (Tables 6 and 7).

Limitations of this review include the retrospective design, the search strategy, databases used,[64] and article quality scoring. The authors conducted a retrospective review with current standards and expectations for scientific rigor that might not have been expected at the time these studies were conducted and published. Authors and indexers are not always on the same page when choosing titles and keywords.[20] Online database searches were inadequate in locating all articles that met the inclusion criteria.[20] Content expert and selective manual searches were necessary in finding many of the articles[20]. The article reviewers each had different education and training backgrounds, accounting for the initial disagreement in scoring in one third of the articles. Before reviewer consensus, there was variability in interpretation of the quality scoring instrument terms as well as in judgments regarding how well an article addressed the issues being evaluated. In using a quality assessment instrument, some quality scoring criteria are more detailed/differentiated than others, which introduces an inherent bias. Scores/assigned weights may be biased toward rigor of research methodology and presentation. Since the

quality assessment instrument focused on the internal validity of the studies, the quality scores cannot be extrapolated to measure the studies' signi cance or impact (in terms of ndings, relevance to the discipline).

There are several strengths, however. The authors formed a multidisciplinary team, paying special attention to minimizing bias by the Doctor of Osteopathic Medicine and Doctor of Chiropractic on our team who did not review studies in their respective professions. The authors combined information (studies) obtained from different professions (PT, DO, DC, MD) in a systematic manner. The quality assessment tool (quality index) was

results of two methods and their relationship to visceral diseaseManual Med 1984;1:25–32.
27. French S, Green S, Forbes A. Reliability of chiropractic methods commonly used to detect manipulable lesions in patients with chronic low-back painJ Manipulative Physiol Ther 2000;23:231–8.
28. Hawk C, Phongphua C, Bleecker J, et al. Preliminary study of the reliability of assessment procedures for indications for chiropractic adjustments of the lumbar spine. J Manipulative Physiol Ther 1999;22:382–9.
29. Jull G, Zlot G, Trott P, et al. Ilb.7812 0 Tcexa-25.7(Jlb.781ncTw [(resE01y4e658.5(detl.)-189na4-26.56ak5(detl.)-0uer)]TJ chronic)-258.5(low-bc:25)Tj /-bc:25Jlb.781ncTw [(r

German][: Zur Zuverlassigkeit manualdiagnostischer Befunde].Manuelle Med 1993;31:1–11.
84. Olson KA, Paris SV, Spohr C, et al. Radiographic assessment and reliability study of the craniovertebral sidebending test.J Manual Manipulative Ther 1998;6:87–96.
85. Lindsay DM, Meeuwisse WH, Mooney ME, et al. Interrater reliability of manual therapy assessment techniques.Physiother Can 1995;47:173–80.
86. Rhudy T, Sandefur M, Burk J. Inter-examiner intertechnique reliability in spinal subluxation assessment: a multifactorial approach.Am J Chiropractic Med 1988;1:111–4.
87. 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.Headache2000;40:581–6.
88. McPartland JM, Goodridge JP. Counterstrain and traditional osteopathic examination of the cervical spine compared.J Bodywork Movement Ther 1997;1:173–8.
89. Hubka MJ, Phelan SP. Inter-examiner reliability of palpation for cervical-spine tenderness.J Manipulative Physiol Ther 1994;17:591–5.
90. Viikari-Juntura E. Inter-examiner reliability of observations in physical examinations of the neck.Phys Ther 1987;67:1526–32.
91. Nice DA, Riddle DL, Lamb RL, et al. Intertester reliability of judgments of the presence of trigger points in patients with low back pain.Arch Phys Med Rehabil 1992;73:893–8.
92. Boline PD, Haas M, Meyer JJ, et al. Inter-examiner reliability of 8 evaluative dimensions of lumbar segmental abnormality. 2.J Manipulative Physiol Ther 1993;16:363–74.
93. Waddell G, Main CJ, Morris EW, et al. Normality and reliability in the clinical assessment of backache.Br Med J (Clin Res Ed) 1982;284:1519–23.
94. Rouwmaat PHM, Everaert D, Stappaerts KH, et al. Reliability of manual skinfold tests in a healthy male population. J Manipulative Physiol Ther 1998;21:327–32.
95. Ghoukassian M, Nicholls B, McLaughlin P. Inter-examiner reliability of the Johnson and Friedman percussion scan of the thoracic spine.J. Osteopath Med 2001;4:15–20.
96. Johnston WL, Allan BR, Hendra JL, et al. Inter-examiner study of palpation in detecting location of spinal segmental