JAOA ORIGINAL CONTRIBUTION

Interobserver Reliability of Osteopathic Palpatory Diagnostic Tests of the Lumbar Spine: Improvements From Consensus Training

Brian F. Degenhardt, DO; Karen T. Snider, DO; Eric J. Snider, DO; and Jane C. Johnson, MA

Context: Establishing reliable palpatory tests continues to be a critical, yet elusive, step in osteopathic medical research and evidence-based clinical practice.

Objective: The authors investigated the interobserver reliability of common osteopathic palpatory tests used to evaluate the lumbar spine.

Design and Methods: Subjects (N=119) were recruited from the faculty, staff, and students of Kirksville (Mo) College of Osteopathic Medicine (KCOM) of A.T. Still University of Health Sciences. Three osteopathic medical examiners residency trained in neuromusculoskeletal medicine initially evaluated lumbar segments on subjects from one subgroup (n=42) in a blinded assessment. The examiners performed palpatory tests of tenderness and tissue texture changes, as well as—in three planes—vertebral positional asymmetry and motion asymmetry. Kappa statistics (κ) were used to evaluate interobserver reliability. Following a period of consensus training, subjects from another subgroup (n=77) were evaluated in a blinded assessment for those palpatory tests that seemed most likely to produce reliable findings. The interobserver reliability was then re-evaluated.

Results: During the initial evaluation of interobserver reliability, κ ranged from −0.02 to 0.34, within the poor-to-fair reliability range. Following consensus training, reliability improved, rising into the moderate range for tissue texture changes (κ=0.45) and into the substantial range for tenderness assessments (κ=0.68). Reliability for positional asymmetry in the transverse plane (κ=0.34) and rotational motion asymmetry (κ=0.20) were improved but remained in the fair range.

Conclusion: The authors concluded that consensus training improved the interobserver reliability of common osteopathic palpatory tests of the lumbar spine. In two of the four tests that were studied—tissue texture and tenderness—acceptable κ values for clinical tests were achieved after consensus training.

From the Kirksville College of Osteopathic Medicine of A.T. Still University of Health Sciences.

Address correspondence to Brian Degenhardt, DO, Director, A.T. Still Research Institute, A.T. Still University of Health Sciences, 800 W Jefferson St, Kirksville, MO 63501–1443.

E-mail: bdegenhardt@atsu.edu

For more than 10 years, evidence-based medicine has challenged osteopathic physicians to use only reliable palpatory tests when diagnosing patients' conditions.[1] Reliability is defined as the reproducibility of findings when a test is repeated to evaluate an unchanged attribute. Because various kinds of palpatory tests are used in patient care within the osteopathic and allopathic medical professions, as well as in chiropractic care and physical therapy, reliability is an important issue for healthcare professionals.

For palpatory tests, two forms of reliability are routinely studied: intraobserver reliability and interobserver reliability. Intraobserver reliability assesses the ability of a healthcare professional to obtain the same finding when serially evaluating a patient. This form of reliability has been criticized as lacking in credibility, mostly because of the difficulties in blinding an examiner between examinations.[2] Interobserver reliability, the degree to which multiple examiners reach the same conclusion, is considered more relevant than intraobserver reliability in assessing practitioner skill.[2] The current study investigated the interobserver reliability of three specialists in osteopathic neuromusculoskeletal medicine.

For at least 30 years, researchers of manual medicine have studied the reliability of many of the commonly used palpatory diagnostic tests.[3] Most of the more than 200 published articles in this area failed to show a level of reproducibility that supports the use of palpation in evidence-based clinical practice. Many of these studies have been criticized for having inadequate research design.[2,4–8] Among the studies that applied rigorous scientific methods, most of these studies reported that, while at least one palpatory test had reasonable reliability, most tests did not.[9–13] Several reviews since the early 1990s have found that because of an insufficient number of well-performed studies, conclusions cannot be arrived at about the reliability of either static landmark position tests or passive motion tests.[8,14,15] As a result, the scientific community continues to question the relevance of manually performed diagnostic tests.

A major criticism of previously performed research is the improper use of statistical methods. Percent agreement, a method used to evaluate reliability of categorical and ordinal ratings in many previous studies, does not account for the amount of agreement that occurs by chance.[8] Thus, percent agreement has been replaced by Cohen's kappa statistic (κ)

Osteopathic Palpatory Diagnostic Test	Description*
■ **Tissue Texture**	The subcutaneous tissue medial and inferior to the transverse processes, down to the level of the facet joints, was palpated for the presence of swelling, "bogginess," and "ropiness."
■ **Transverse Plane**	
□ Positional asymmetry	The posterior surfaces of the transverse processes of each vertebral segment were palpated. Deviation from the neutral position was determined.
□ Motion asymmetry	*Vertebral rotation*—Anterior pressure was applied in an alternating manner to the posterior tips of the transverse processes to determine motion asymmetry in the transverse plane.
	Anterior-posterior translational springing†—Thumb tips or hypothenar eminences were placed on the vertebral spinous process. Pressure was applied anteriorly in a springing manner to determine freedom of motion.
■ **Coronal Plane**	
□ Positional asymmetry	The inferior surfaces of the vertebra's transverse processes were palpated. The relative position was compared to the neutral position.
□ Motion asymmetry	Translatory pressure was placed between vertebral units, below the medial portion of the transverse processes, to determine motion asymmetry in the coronal plane.
■ **Sagittal Plane**	
□ Positional asymmetry	The spinous process of each segment was compared with neighboring spinous processes. Cephalad/caudad asymmetry was determined.
□ Motion asymmetry	*Flexion preference*—With subjects bent forward in the seated position, examiners applied uniform anterior force on both transverse processes of each vertebral segment to determine flexion preference.
	Extension preference—Subjects were prone with elbows resting on table to cause back to arch. Pressure was applied to the transverse processes of a single vertebral segment to determine extension preference.
■ **Tenderness**	With subjects prone, anteriorly directed force was applied on each spinous process from L1–L4, with pressure gradually increasing from 0 kg/cm^2 to at least 4 kg/cm^2. The examiners then reproduced the amount of pressure applied on a calibrated scale next to subjects to quantify the amount of pressure that induced pain.

* Unless otherwise noted, all palpatory diagnostic tests were performed with the subjects lying in the prone position.
† Examiners investigated the palpatory anterior-posterior translational springing test for motion asymmetry in phases 2 and 3 only.

Figure 1. *Osteopathic palpatory diagnostic tests that examiners performed on subjects' lumbar vertebrae (L1–L4) to test interobserver reliability for the current study.*

as the preferred measure of the strength of reliability.

Kappa statistic provides for a more stringent test because it accounts for not only the agreement among observers but also the prevalence and variability of the observations. If there is a preponderance of a particular finding, the percent agreement may increase, but the κ value may decrease because of lack of observation variability. Landis and Koch[16] established a scale for interpreting κ values as follows: 0.81–1.00 indicates almost perfect reliability; 0.61–0.80, substantial reliability; 0.41–0.60, moderate reliability; 0.21–0.40, fair reliability; and less than 0.20, poor reliability. When analyzing the results of subjects' physical examinations, a κ value of at least 0.40 is considered an indicator of acceptable interobserver reliability.[17]

There are four general types of palpatory diagnostic tests commonly taught and used in clinical practice, including osteopathic medicine. Clinicians use these tests to perform the following evaluations:

□ Differentiation of tissue textures,
□ evaluation of static landmark positional asymmetry,

- evaluation of motion asymmetry, and
- assessment of tenderness.[18–20]

In a systematic review, Hestbaek and Leboeuf-Yde[15] reported that palpatory tests assessing tenderness have consistently shown at least moderate interobserver reliability. However, palpatory tests of landmark positional asymmetry, motion asymmetry, and tissue texture have consistently shown poor reliability.[5,11,13,15,21–23]

We propose that acceptable interobserver reliability was not found in earlier studies of palpatory diagnostic tests for several reasons. First, human beings are not static entities. Homeostatic mechanisms, such as heart rate, respiratory rate, blood pressure, and—especially pertinent for the current study—neuromotor reflexes are responsible for constant inherent neurophysiologic variability. The neuromusculoskeletal system changes on some level each second according to the impulses or stresses that individuals experience. This inherent neurophysiologic variability occurs in both the examiner and subject. The dynamic nature of the human body could thus challenge clinicians' abilities to reliably perform palpation because, by definition, reliability determines the reproducibility of findings when a test is repeated to evaluate an unchanged attribute. Second, some current educational systems may not provide students or clinicians with the necessary skills for reliably performing palpation. Third, because of the relative isolation of many private practices, experts in manual medicine may not perform palpatory procedures in similar ways, limiting the likelihood of reliability.

We hypothesized that examiners would have improved interobserver reliability after participating in a rigorous consensus-training program. Because palpation that induced repetitive motion would be more likely than positional palpation to change the characteristics being examined, we also predicted that positional asymmetry tests would require less training than motion asymmetry testing to obtain at least moderate interobserver reliability. In addition, this consideration led us to predict that κ values for tests that induce motion would be lower than κ values for positional asymmetry tests.

Methods

The institutional review board at Kirksville (Mo) College of Osteopathic Medicine (KCOM) of A.T. Still University of Health Sciences approved the protocol of the current study prior to subject recruitment. The subjects in this study were healthy volunteers recruited from KCOM faculty, staff, and students. All of the subjects signed informed consent forms approved by the college's institutional review board.

The study was designed to determine which palpatory diagnostic tests of the lumbar spine demonstrated at least moderate interobserver reliability when performed by three examiners with osteopathic residency training in neuromusculoskeletal medicine. Most of the tests and the parameters used to interpret them were well established within the osteopathic medical profession.[18–20] For manual tests that failed to show at least moderate interobserver reliability using the skills developed through basic and advanced osteopathic manipulative training, the examiners underwent consensus training to improve reliability prior to re-evaluating the tests.

Following test identification and definition, the experimental design consisted of the following three phases:

- Phase 1: Pretraining interobserver reliability assessment;
- Phase 2: Consensus training; and
- Phase 3: Posttraining interobserver reliability assessment.

Three examiners (B.F.D.,K.T.S.,E.J.S.), who are members of KCOM's Department of Osteopathic Medicine, participated in this study. All three examiners had residency training in neuromusculoskeletal medicine and fewer than 10 years clinical experience in the specialty. A single-blinded method—with examiners not blinded—was used for test identification\definition and for phase 2 of this study. The examiners collected data in a double-blinded method in phases 1 and 3. Only data collected in phases 1 and 3 were used in the statistical analyses.

Test Identification and Definition

In the test identification and definition part of the current study, examiners determined which palpatory diagnostic tests commonly used to evaluate lumbar segments L1 through L4 should be included in the study. Evaluation of the L5 vertebral segment was not included in the study because of the significant anatomic variability in size, shape, and relative positions of L5 and the ilium and sacrum.

The examiners reviewed standardized descriptions of the tests (*Figure 1*) and parameters for interpreting those tests. The tests consisted of assessments of tenderness and tissue texture changes, as well as evaluations of positional and motion asymmetry in three planes (coronal, sagittal, and transverse).

Although preferred methods for most of the palpatory tests have long been agreed upon in the osteopathic medical profession,[18–20] tests for positional asymmetry in the coronal plane and motion asymmetry in the sagittal plane are not well established. Therefore, the examiners agreed upon methods for these tests that appeared consistent with the models used to test vertebral motion in other reference planes.

To test for positional asymmetry in the coronal plane with the subject in the prone position, thumbs were placed in the interspaces between the transverse processes of two adjacent vertebrae. Gradual cephalad motion was performed until a firm barrier was noted to indicate that the thumbs were on the under surfaces of the transverse processes. The diagnosis was based on the relative position of the transverse processes compared to the neutral position (ie, right thumb inferior for side-bending to the right). To test for flexion motion preference in the sagittal plane, seated subjects were bent forward, and anterior springing motion was induced on both transverse processes simultaneously. To determine extension motion

preference, prone patients arched their backs and rested on their elbows in a sphinxlike position. Examiners used their thumbs to apply pressure on the transverse processes, augmenting extension.

Phase 1: Pretraining Interobserver Reliability Assessment

Forty-two healthy subjects between the ages of 20 and 44 years (mean age, 26 years [SD±4]) were recruited for phase 1 of the current study from the faculty, staff, and students of KCOM. Thirty-one (74%) men and 11 (26%) women participated in phase 1 of the current study.

Subjects were placed in 14 groups of three subjects. The three examiners worked simultaneously in each training session, providing physical examinations to one group of subjects per session. These sessions took a total of approximately five hours and were spread over four days. This phase of the study was conducted in a double-blinded manner, with each examiner blinded to the findings of the other two examiners.

Examiners asked subjects either to lie in the prone position (for most tests) or bend forward in the seated position (for some tests of motion asymmetry in the sagittal plane), depending on the test being performed, and to refrain from extraneous movements until the physical examination was complete. Each of the three examiners performed all eight palpatory tests on one subject in each group and recorded the presence or absence of physical findings after each examination. Each examiner repeated palpatory tests up to three times until that examiner determined a consistent finding.

The results of each test were analyzed based on variants of κ statistics for two and three examiners, corresponding 95% confidence intervals (CI), and percent agreement. The prevalence of the findings in the sample was calculated as a weighted proportion in which the weights were based on the number of examiners indicating vertebrae were positive for a given palpatory finding. The results were reviewed, and those tests with a κ value of less than 0.40 were considered for further study in phases 2 and 3.

Phase 2: Consensus Training

For one to two hours a week over a four-month period, the three examiners performed a protocol designed to promote consensus, focusing on one test at a time. Initially, examiners evaluated subjects simultaneously, observing each other's testing procedures. The subjects remained motionless in the prone position during these examinations.

When each examiner expressed confidence that adequate modifications had been made to his or her technique to improve reliability, the examiners performed blinded examinations on a limited number of subjects to determine if progress had been made. Once a high percent agreement was identified for a particular test, consensus training was continued with the next test.

Phase 3: Posttraining Interobserver Reliability Assessment

For each of the tests used in phase 2, interobserver reliability was reassessed using 15 to 33 new subjects. The precise number of subjects used depended on the testing method under investigation; more subjects were used if the examiners believed a test needed more work to achieve acceptable reliability. All these new subjects came from a group of 77 individuals, also healthy and between the ages of 20 and 65 (mean age, 31 [SD±8]), who were also recruited from KCOM's faculty, staff, and students. Thirty (39%) men and 47 (61%) women participated in phase 3 of the study.

Consistent with the protocol used in phase 1, groups of three subjects were examined simultaneously, each examiner evaluating one subject at a time in a blinded manner. Once an examination was completed, the examiners switched positions until all subjects had been examined. The subjects remained still in the prone position throughout testing.

When examiners performed palpatory tests assessing positional asymmetry and tissue texture on the subjects in a group, they always tested positional asymmetry first. The examiners performed the other two tests on separate days. Because of scheduling conflicts, only two of the three examiners (B.F.D.,K.T.S.) performed the test for tenderness. In these circumstances, only two subjects were examined at a time. Phase 3 testing took seven hours, which were spread over a six-day period.

After completing tests on each set of three subjects, the examiners scanned the subjects' data forms, identifying cases in which examiners had diagnosed opposite findings. If opposite findings were diagnosed for a vertebra, each examiner would re-evaluate that specific vertebra while the other examiners closely watched how the test was being performed. Through observation and subsequent discussion, a reason for each diagnostic variation was determined, and the testing procedures were modified if necessary.

As noted, data in phase 3 were collected in a double-blinded method and analyzed based on variants of κ for two and three examiners, corresponding 95% CI, and percent agreement. Likewise, the prevalence of the findings in the sample was calculated as a weighted proportion in which the weights were based on the number of examiners who indicated the vertebrae were positive for given palpatory findings. Logistic regression models were used to test for change in the probability of agreement between examiners from preconsensus to postconsensus training.

Results

In phase 1 of the study, tenderness demonstrated fair interobserver reliability (κ=0.34). All other palpatory tests studied demonstrated slight to poor interobserver reliability (κ<0.20) (*Figure 2*). Tenderness, tissue texture changes, and transverse plane tests (motion asymmetry and positional asymmetry)

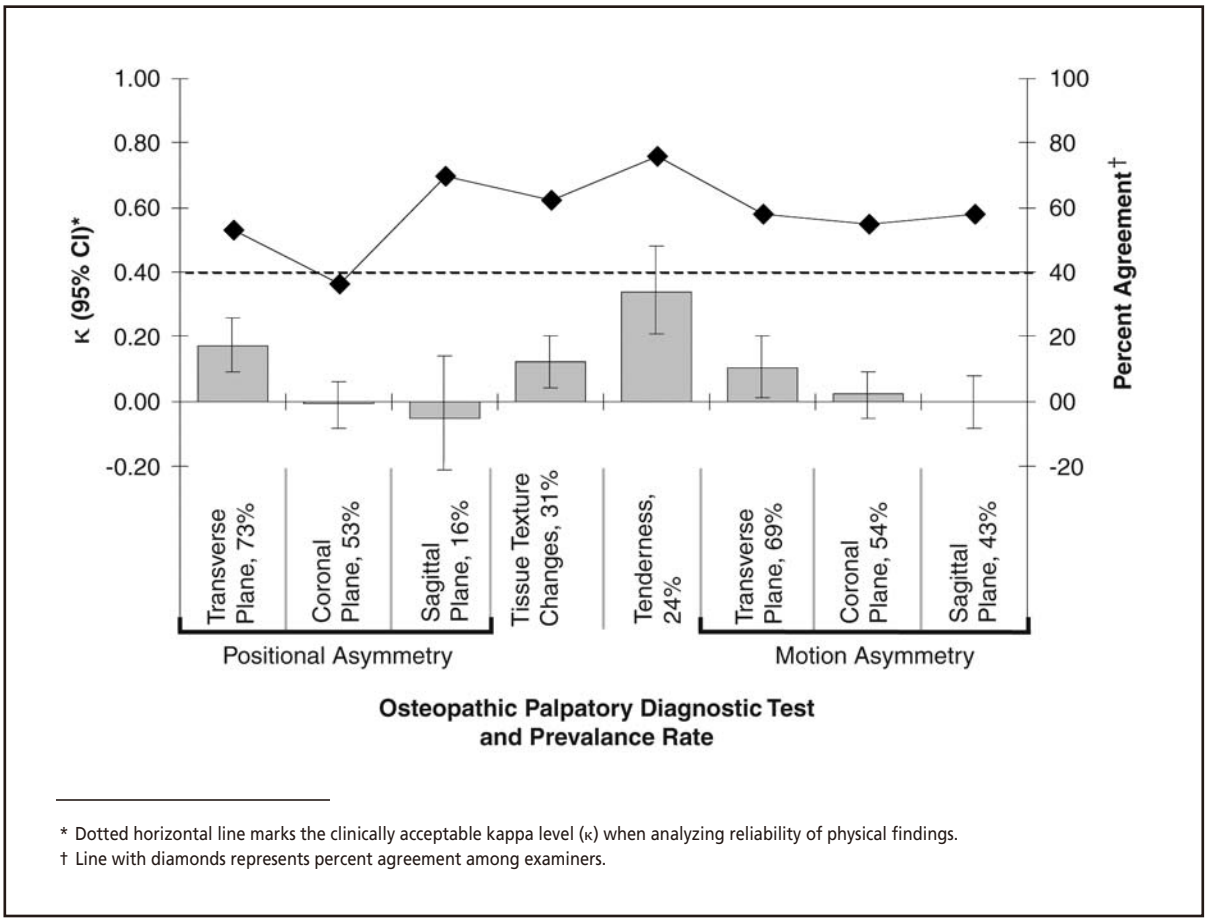


Figure 2. *In the preconsensus training reliability assessment (phase 1), osteopathic palpatory diagnostic tests for tissue texture changes, tenderness, and positional and motion asymmetry in the transverse, coronal, and sagittal planes were evaluated in subjects (n=42). Tissue texture changes, tenderness, and transverse plane tests showed the greatest likelihood for consensus training to improve interobserver reliability, as expressed by kappa statistic (κ).*

had the highest κ values, indicating the greatest likelihood for improvement of interobserver reliability with consensus training.

In phase 2, consensus training was performed for those palpatory tests with the highest interobserver reliability. Tests for static landmark position asymmetry, tissue texture changes, and tenderness together took a total of three sessions of training until the examiners felt confident that their interobserver reliability values had adequately improved. These sessions began with the examiners simultaneously evaluating one subject. The examiners observed each other's performance of the test and then discussed observed variations in their techniques. This process allowed for the examiners to refine and perform their techniques in a standard manner.

Once the examiners were confident that their technique had been standardized, blinded examinations were performed on one subject. Immediately afterward, the findings of the different examiners were compared. For each vertebra that was a subject of disagreement, each member of the team repeated the test with the other examiners observing. Further discussion ensued to further refine technique and establish consensus.

The examiners spent 12 sessions over three months trying to establish reliability on the test assessing rotational motion. However, despite single-blinded interactions, interobserver reliability did not improve for this test.

To better understand the mechanics of each examiner's testing style, one examiner's thumbs rested passively on the tips of a subject's transverse processes while another examiner performed motion testing through the thumbs of the first examiner. Although initially useful, continued experimentation suggested that this training approach was unlikely to prove significantly beneficial. In recognition of the importance of motion asymmetry testing to osteopathic medicine, the motion testing protocol was changed to an anterior-posterior translational springing of the lumbar spinous processes, which proved to be of greater benefit during training.

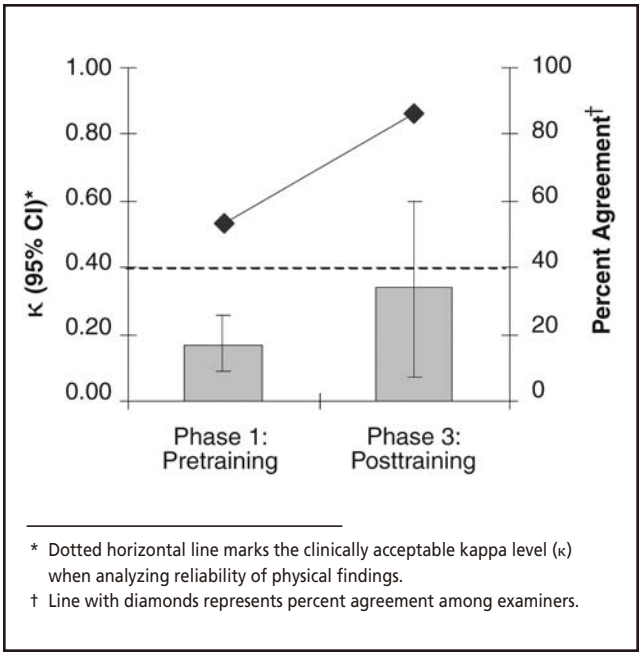


Figure 3. *The osteopathic palpatory diagnostic test for positional asymmetry in the transverse plane (P<.001) failed to reach the clinically acceptable reliability range in the postconsensus training reliability assessment (phase 3), in which interobserver reliability was expressed by kappa statistic (κ). The P values were calculated using logistic regression models to test for change in the probability of agreement among examiners from preconsensus to postconsensus training. Prevalence in the pretraining group of subjects (n=42) for positional asymmetry changes was 73%. Prevalence in the posttraining group (n=30) was 94%.*

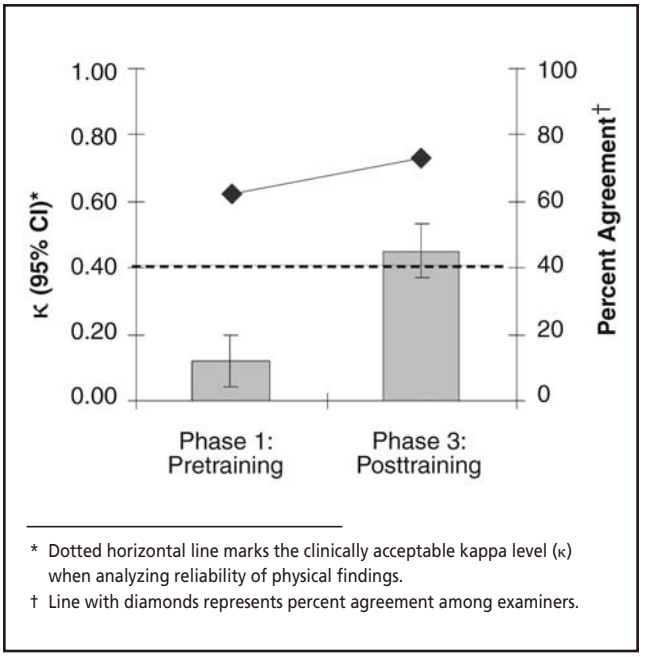


Figure 4. *The osteopathic palpatory diagnostic test for tissue texture (P=.003) reached the moderate reliability range in the postconsensus training reliability assessment (phase 3), in which interobserver reliability was expressed by kappa statistic (κ). The P values were calculated using logistic regression models to test for change in the probability of agreement among examiners from preconsensus to postconsensus training. Prevalence in the pretraining group of subjects (n=42) for tissue texture changes was 31%. Prevalence in the posttraining group (n=30) was 59%.*

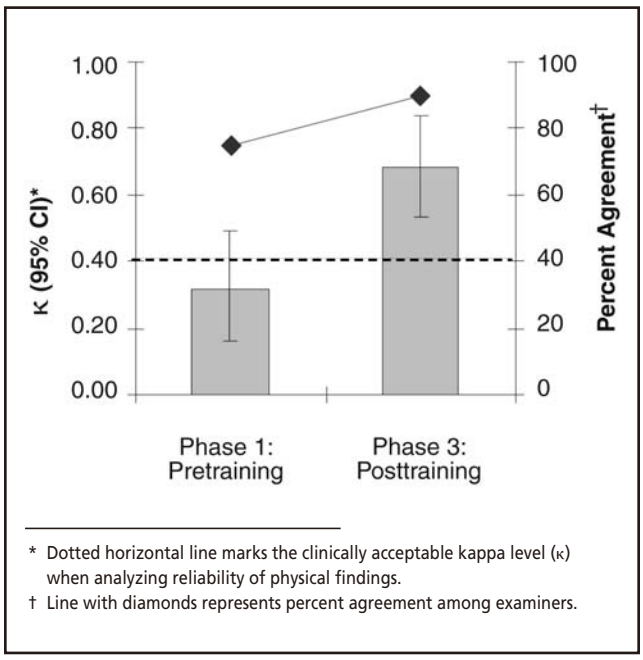
Results from phase 3 of the study (*Figures 3–6*) indicated significant improvement in interobserver reliability, into the clinically acceptable range (κ≥0.40), for tests of tenderness (preconsensus training κ=0.32, postconsensus training κ=0.68; *P*=.02) and tissue texture (preconsensus training κ=0.12, postconsensus training κ=0.45; *P*=.003). Although interobserver reliability was significantly improved for positional asymmetry in the transverse plane (preconsensus training κ=0.17, postconsensus training κ=0.34; *P*<.001), adequate interobserver reliability was not obtained (κ=0.40). By adapting the test for motion asymmetry in the transverse plane to anterior-posterior translational springing of the lumbar spinous processes, interobserver reliability improved (preconsensus training κ=0.10, postconsensus training κ=0.20; *P*=.04), but it did not reach an adequate level.

Comment

The current study demonstrates that it is possible to significantly improve the interobserver reliability of palpatory tests used by osteopathic physicians to diagnose somatic dysfunction of the lumbar vertebrae. Three clinicians residency-trained in neuromusculoskeletal medicine were able to improve their interobserver reliability on four palpatory tests after they established a consistent framework in which to perform and interpret those tests.

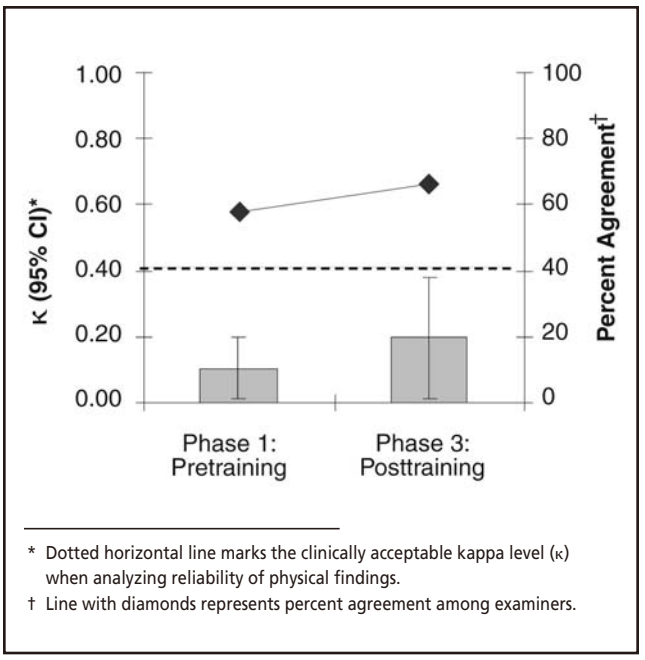
For two of these palpatory tests (tissue texture changes and tenderness assessment), interobserver reliability improved from fair or poor reliability to substantial or moderate reliability, respectively. The other two tests (assessing vertebral positional asymmetry and motion asymmetry) showed improvement from poor to fair reliability.

A qualification in the interpretation of the results of this study must be noted, however. The demographic characteristics of the subject pool may have influenced the results of the study. There were age and gender differences between the pretraining and posttraining groups. For example, in the pretraining group (phase 1), mean subject age was 26, and 74% of the subjects were men. In the posttraining group (phase 3),



* Dotted horizontal line marks the clinically acceptable kappa level (κ) when analyzing reliability of physical findings.
† Line with diamonds represents percent agreement among examiners.

Figure 5. *The osteopathic palpatory diagnostic test for tenderness (P=.02) reached the substantial reliability range in the postconsensus training reliability assessment (phase 3), in which interobserver reliability was expressed by kappa statistic (κ). The* P *values were calculated using logistic regression models to test for change in the probability of agreement among examiners from preconsensus to postconsensus training. Prevalence in the pretraining group of subjects (n=42) for tenderness changes was 24%. Prevalence in the posttraining group (n=33) was 18%.*



* Dotted horizontal line marks the clinically acceptable kappa level (κ) when analyzing reliability of physical findings.
† Line with diamonds represents percent agreement among examiners.

Figure 6. *The osteopathic palpatory diagnostic test for motion asymmetry in the transverse plane (P=.04) failed to reach the clinically acceptable reliability range in the postconsensus training reliability assessment (phase 3), in which interobserver reliability was expressed by kappa statistic (κ). The* P *values were calculated using logistic regression models to test for change in the probability of agreement among examiners from preconsensus to postconsensus training. Prevalence in the pretraining group of subjects (n=42) for motion asymmetry was 69%. Prevalence in the posttraining group (n=15) was 76%.*

mean subject age was 31, and 39% of the subjects were men. In addition, the subjects' body mass indices, which could have played a role in the results, were not recorded. Thus, it is possible that the interobserver reliability changes observed in the current study were secondary to these demographic and physical differences, rather than being the result of consensus training. In future studies, investigators may wish to use the same group of subjects in both the preconsensus and postconsensus trials.

These cautions are somewhat mitigated by the fact that all the subjects (N=119) were young (mean age, 29 years [SD±7]). Such subjects would be expected to have less severe and less frequent somatic dysfunctions compared with typical patients of osteopathic family physicians. By studying younger and healthier subjects than those who seek medical intervention, the examiners in this study may have evaluated individuals who have less somatic dysfunction and possibly more reactive neuromusculoskeletal reflexes than the typical patient. As a result, palpatory findings may have changed during the repetitive diagnostic process. If these assumptions are true, the κ values determined in this study would be lower than what would probably occur in the typical clinic setting.

It would be beneficial for further investigations to examine subjects throughout all stages of life, evaluating palpatory tests for all relevant regions of the body and enrolling a significant number of individuals who are symptomatic and have clinically meaningful dysfunction. Such a study design is likely to produce a more conclusive definition or demonstration of the reliability of osteopathic palpatory testing methods.

The poor κ scores obtained in phase 1 may have resulted from at least two confounding factors: inappropriate methodology and limitations in the training of the examiners. During phase 1, each examiner performed eight tests on one subject before moving to the next subject. Each test was performed two or three times until the individual examiner was confident of that test's outcome. As a result, each vertebra was tested between 48 and 72 times. Thus, changes in the subjects' neu-

romusculoskeletal systems secondary to repetitive stimuli may have been one reason for poor reliability in phase 1 of the study.

Another methodological issue that arose during phase 1 is that the subjects had to change position for each examiner so that all the tests could be performed sequentially. Consequently, variations in subject position from the first examiner to the two subsequent examiners could have been an additional confounding factor. Examiners addressed both of these issues in phases 2 and 3 of the study by requesting that subjects remain in the same prone position throughout the testing sequence. In addition, in phases 2 and 3, each vertebra was tested no more than 18 times by the examiners.

In phase 2, the consensus training for nonmotion palpatory diagnostic tests consisted of only one training session per test. However, the test that induced vertebral motion in the transverse plane did not significantly improve in reliability even after 12 sessions of training. The examiners noted that motion characteristics routinely changed between one examiner's evaluation and the next. This experience supports the hypothesis that motion testing, which stimulates the sensory nervous system, may cause neuromotor reflexes to adapt to the stimuli—especially in young individuals like this study's subjects—causing findings to change after repetitive stimuli.

Also during phase 2, it was obvious that there were unique aspects of each examiner's technique. This variety would seem to indicate that the skills the examiners developed in their osteopathic medical educational programs or medical practices were not—or were no longer—standardized.

The current study's research team excelled in palpatory skills during their undergraduate and graduate osteopathic medical education. All three examiners trained in the same two-year residency program in neuromusculoskeletal medicine and osteopathic manipulative medicine. Two had completed the program, and the third was in the final months of completing the program while participating in the current study. The ability to apply osteopathic palpatory diagnostic skills in a reliable, scientific format is most likely to exist among those physicians who have the greatest and most similar training in this area (board-certified and board-eligible, residency-trained osteopathic physicians). Nevertheless, the palpatory experts in the current study were initially unable to demonstrate moderate reliability of any of the tests performed. This inability was consistent with numerous previous studies in the literature.[2–8]

To address this problem and to better calibrate and standardize the palpatory skills of future osteopathic physicians, osteopathic medical educators need to critically evaluate the current methods by which palpatory diagnostic skills are taught and reinforced in schools. The current study suggests that to date, the curricula at osteopathic medical schools have not been formatted to determine the interobserver reliability skills of each graduating student. Although the challenges in formatting and the logistics in developing such curricula are understood, osteopathic medical educators need to appreciate the importance of overcoming these obstacles. Testing formats need to be revised to ensure that the baseline skills in palpatory diagnostics for all osteopathic medical students have a level of reliability that would be acceptable within the scientific community ($\kappa \geq 0.40$).

One process for establishing improved training methods would be to identify those palpatory diagnostic tests that are most reliable in the hands of clinical researchers. Then those methods should be used as standard procedures in the teaching and testing of palpatory skills in osteopathic medical schools. Once adequate interobserver reliability for a test has been demonstrated by a group of researchers, clinicians, or students, the persistence of the test's interobserver reliability should be determined to establish whether that palpatory instrument can remain calibrated.

Although it is fundamental to the scientific process to establish reliable palpatory diagnostic tests, the significance of establishing appropriate levels of interobserver reliability is limited. Just because any number of clinicians can demonstrate moderate to substantial interobserver reliability of certain tests, one cannot generalize that these tests are reliable in the hands of all osteopathic physicians who perform them. Even if the osteopathic medical profession's colleges, internships, and residencies could establish programs that demonstrate the reliability of the performance of their graduates, reliability alone does not address issues concerning the accuracy, validity, and clinical impact of the palpatory tests used by these graduates. These are crucial issues that must be more aggressively studied.

Conclusion

Although initially unsuccessful with eight commonly used osteopathic palpatory tests, three examiners residency-trained in neuromusculoskeletal medicine improved with consensus training the interobserver reliability values of osteopathic palpatory diagnostic tests assessing tissue texture changes, tenderness, and positional and motion asymmetry in the transverse plane. This success demonstrated that consensus training can be effective in significantly improving the interobserver reliability of palpatory tests. However, additional research is needed to address whether the demographic characteristics of subject pools may influence the results. Furthermore, the results of the current study suggest that osteopathic medical educators need to modify their curricula to better calibrate and standardize palpatory diagnostic skills.

Acknowledgments

The authors thank Rori Caruthers for her technical support.

This study was supported by the National Institutes of Health's National Center for Complimentary and Alternative Medicine (Grant No. 1R01AT00305-1) and the American Osteopathic Association (Grant No. 00-04-505).

References

1. Library of the Health Sciences–Peoria. Evidence based medicine: finding the best clinical literature. The University of Illinois at Chicago Web site. April 1, 2005. Available at: http://www.uic.edu/depts/lib/lhsp/resources/ebm.shtml. Accessed August 5, 2005.

2. Haas M. The reliability of reliability [review]. *J Manipulative Physiol Ther*. 1991;14:199–208.

3. Johnston WL. Interexaminer reliability studies: spanning a gap in medical research—Louisa Burns Memorial Lecture. *J Am Osteopath Assoc*. 1982;81:819–829.

4. Alley JR. The clinical value of motion palpation as a diagnostic tool: a review. *J Can Chiropr Assoc*. 1983;27:97–100.

5. Russell R. Diagnostic palpation of the spine: a review of procedures and assessment of their reliability. *J Manipulative Physiol Ther*. 1983;6:181–183.

6. Keating JC Jr. Several strategies for evaluating the objectivity of measurements in clinical research and practice. *J Can Chiropr Assoc*. 1988;32:133–138.

7. Keating JC Jr. Inter-examiner reliability of motion palpation of the lumbar spine: a review of the quantitative literature [review]. *Am J Chiropractic Med*. 1989;2:107–110.

8. Haas M. Statistical methodology for reliability studies. *J Manipulative Physiol Ther*. 1991;14:119–132.

9. Mior SA, King RS, McGregor M, Bernard M. Intra- and interexaminer reliability of motion palpation of the cervical spine. *J Can Chiropr Assoc*. 1985;29:195–198.

10. Boline PD, Keating JC Jr, Brist J, Denver G. Interexaminer reliability of palpatory evaluations of the lumbar spine. *Am J Chiropractic Med*. 1988;1:5–11.

11. Mootz RD, Keating JC Jr, Kontz HP, Milus TB, Jacobs GE. Intra- and interexaminer reliability of passive motion palpation of the lumbar spine. *J Manipulative Physiol Ther*. 1989;12:440–445.

12. Keating JC Jr, Bergmann TF, Jacobs GE, Finer BA, Larson K. Interexaminer reliability of eight evaluative dimensions of lumbar segmental abnormality. *J Manipulative Physiol Ther*. 1990;13:463–470.

13. Jull G, Bogduk N, Marsland A. The accuracy of manual diagnosis for cervical zygapophysial joint pain syndromes. *Med J Aust*. 1988;148:233–236.

14. Panzer DM. The reliability of lumbar motion palpation [review]. *J Manipulative Physiol Ther*. 1992;15:518–524.

15. Hestbaek L, Leboeuf-Yde C. Are chiropractic tests for the lumbo-pelvic spine reliable and valid? A systematic critical literature review [review]. *J Manipulative Physiol Ther*. 2000;23:258–275.

16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.

17. Fjellner A, Bexander C, Faleij R, Strender LE. Interexaminer reliability in physical examination of the cervical spine. *J Manipulative Physiol Ther*. 1999;22:511–516.

18. Dinnar U, Beal MC, Goodridge JP, Johnston WL, Karni Z, Mitchell FL Jr, et al. Classification of diagnostic tests used with osteopathic manipulation. *J Am Osteopath Assoc*. 1980;79:451–455.

19. Dinnar U, Beal MC, Goodridge JP, Johnston WL, Karni Z, Mitchell FL Jr, et al. Description of fifty diagnostic tests used with osteopathic manipulation. *J Am Osteopath Assoc*. 1982;81:314–321.

20. Kuchera WA, Kappler RE. Musculoskeletal examination of somatic dysfunction. In: Ward RC, ed. *Foundations for Osteopathic Medicine*. 2nd ed. Philadelphia, Pa: Lippincott Williams and Wilkins; 2002:633–659.

21. Love RM, Brodeur RR. Inter- and intra-examiner reliability of motion palpation for the thoracolumbar spine. *J Manipulative Physiol Ther*. 1987;10:1–4.

22. Spring F, Gibbons P, Tehan P. Intra-examiner and inter-examiner reliability of a positional diagnostic screen for the lumbar spine. *J Osteopath Med (Australia)*. 2001;4:47–55.

23. French SD, Green S, Forbes A. Reliability of chiropractic methods commonly used to detect manipulable lesions in patients with chronic low-back pain. *J Manipulative Physiol Ther*. 2000;23:231–238.