0008-3194/2013/156–164/$2.00/©JCCA 2013

Interexaminer reliability of cervical motion palpation using continuous measures and rater confidence levels

Robert Cooperstein, MA, DC*
Morgan Young, DC*
Michael Haneline, DC, MPH**

Introduction: *Motion palpators usually rate the movement of each spinal level palpated, and their reliability is assessed based upon discrete paired observations. We hypothesized that asking motion palpators to identify the most fixated cervical spinal level to allow calculating reliability at the group level might be a useful alternative approach.*

Methods: *Three examiners palpated 29 asymptomatic supine participants for cervical joint hypomobility. The location of identified hypomobile sites was based on their distance from the T1 spinous process. Interexaminer concordance was estimated by calculating Intraclass Correlation Coefficient (ICC) and mean absolute differences (MAD) values, stratified by degree of examiner confidence.*

Results: *For the entire participant pool, ICC [2,1] = 0.61, judged "good." MAD=1.35 cm, corresponding to mean interexaminer differences of about 75% of one cervical vertebral level. Stratification by examiner confidence levels resulted in small subgroups with equivocal results.*

Discussion and Conclusion: *A continuous measures*

Introduction *: Généralement, la palpation évalue le mouvement de chaque niveau de la moelle épinière palpé, et sa fiabilité est évaluée sur des observations jumelées et séparées. Nous avons émis l'hypothèse que l'utilisation de la palpation afin d'identifier le niveau de la moelle épinière cervicale le moins mobile, dans le but de permettre le calcul de sa fiabilité à l'échelle du groupe pourrait être une approche alternative utile.*

Méthodologie *: Trois examinateurs ont palpé 29 participants asymptomatiques allongés atteints d'hypomobilité de l'articulation cervicale. L'emplacement de ces segments hypomobiles s'est fondé sur leurs distances par rapport à l'apophyse épineuse T1. La concordance entre les examinateurs a été estimée en calculant le coefficient de corrélation interne (ICC) et les valeurs de la différence absolue moyenne (MAD), stratifiés selon le degré de confiance de l'examinateur.*

Résultats : *Pour tout le bassin de participants, ICC [2,1] = 0,61, jugé « bon ». MAD = 1,35 cm, ce qui correspond à la différence moyenne entre les examinateurs d'environ 75 % d'un segment de la colonne vertébrale. La stratification par le niveau de confiance de l'examinateur a entraîné des petits sous-groupes avec des résultats équivoques.*

Discussion et conclusion : *Une méthodologie ayant recours à des mesures continues pour l'évaluation de*

* Palmer Center for Chiropractic Research, San Jose, CA, USA
** International Medical University, Kuala Lumpur, Malaysia
Research conducted at Palmer College of Chiropractic; San Jose, CA, USA
Responsible author:
Robert Cooperstein, Palmer Chiropractic College, 90 East Tasman Drive, San Jose CA 95134
408 944 6009 voice
408 944 6118 fax
Work supported exclusively by Palmer Center for Chiropractic Research
Protection of human participants
The methods used in this study conformed to the ethical standards of the institutional review board of the college and to the Helsinki Declaration.

study methodology for assessing cervical motion palpation reliability showed more examiner concordance than was usually the case in previous studies using discrete methodology.

KEY WORDS: motion palpation, fixation, cervical spine, concordance

la fiabilité de la palpation de la colonne vertébrale a indiqué une plus grande concordance entre les examinateurs qu'à l'accoutumée lors des précédentes études, qui utilisaient la méthodologie séparée.

MOTS CLÉS : palpation, focalisation, colonne vertébrale, 1concordance

Introduction

Motion palpation (MP) in one form or another is integral to most chiropractic techniques, and is found within the core curriculum at virtually every institution where manual therapy procedures are taught and practiced. Given its ubiquity and strategic importance in training programs, the intraexaminer and interexaminer reliability of MP have been extensively studied and summarized in systematic and annotated reviews.[1-5] In their review of 44 MP studies, Haneline et al[6] reported that only 8 showed high levels of reliability, and that only 2 of these 8 studies could be judged to be of high quality. MP has been found so unreliable[7-9] that some have controversially called for abandoning this diagnostic procedure[10], while Hestbaek and Leboeuf-Yde concluded "The esteem chiropractors have for motion palpation in particular has not been substantiated by scientific data"[11].

Cooperstein et al[12] hypothesized that the design methods of previous interexaminer MP reliability studies may not have been optimal to evaluate interexaminer agreement. All such studies, despite some differences, shared the method of analyzing agreement on a segmental basis. That is, the examiners tested and compared impressions for each spinal level considered separately.

Some of the earlier studies reported the results in terms of percentage agreement, but as Haas pointed out[13], this does not correct for chance agreement. All the more recent studies have assessed concordance using the kappa statistic, which does indeed correct for chance agreement. Sim and Wright have described several factors that can influence the magnitude of kappa, including prevalence and bias, and discussed ways of interpreting the magnitude of obtained kappa values.[14]

Assessing agreement level by level may not reflect the conceptual model that some doctors in clinical practice use when asked to compare opinions for a specific patient on the location of hypomobility. For example, asked to evaluate the cervical spine of a patient with a "stiff neck," we suspect some doctors would attempt to identify the most hypomobile level in the neck. The levels they found could then be judged to be either *relatively close to or distant* from one another. Then, we could conclude the doctors had closely agreed upon, almost agreed upon, or simply disagreed about the location of the most hypomobile segment. Cooperstein et al[12] studied the interexaminer reliability of thoracic MP using this conceptual model and the statistical method best adapted to this type of analysis, the Intraclass Correlation Coefficient (ICC). Although the kappa statistic performs calculations on discrete paired observations, ICC performs calculations on continuous data at the group level.

Cooperstein et al[12] also reasoned that many participants in previous MP studies (often done using largely asymptomatic students) may have lacked a significant hypomobile location, forcing examiners to opine "fixated" or "not-fixated" at each level including cases where they were simply not sure of their findings. To take this into account, examiners in their thoracic study were asked to rate their confidence in the finding each time they palpated a participant. Then, in analyzing the results, examiner agreement could be calculated among several subsets of study participants, stratified by the degree of doctor confidence. Without stratification by doctor confidence, interexaminer reliability was "poor": ICC[2,1] = .3110 (95% CI, 0.0458, 0.5358). In contrast, when both examiners were very confident, interexaminer agreement was "excellent": ICC[2,1] =0.8266 (95% CI, 0.6257, 0.9253). The objective of our present study was to apply Cooperstein's thoracic spine methodology to the cervical spine, for which, to our knowledge, no data using continuous analysis and stratified confidence ratings have been previously reported.

Methods

This studied required and received approval from the Institutional Review Board at our college. All participants were required to provide informed written consent prior to being enrolled in the study. The participants were a convenience sample of asymptomatic chiropractic students who volunteered to participate during a technique laboratory class. Participants with reported cervical pain greater than 2/10 or intolerance to the palpation procedure for any reason were excluded. There were no other exclusion criteria. Participants (n=29) were mostly male (n=19), mean 27.1 years of age, mean weight 71.2 kg, and mean height 172.3 cm. The mean pain level was 0.8 on an 11-point numeric pain scale, as established by participant-completed questionnaires. No potential participants were excluded.

The 3 examiners used in this study were licensed chiropractors, two with more than 20 years of clinical experience and one with approximately 3 years of experience. The participants were instructed not to speak to the examiners during the examination process and were unaware of the palpatory results. The sequence of examiners was randomized for each participant by means of an (unblinded) research assistant drawing color-coded slips of paper from an envelope to prevent order effects. Each examiner was masked as to other examiners' findings.

Participants were first placed prone in order to permit a research assistant to mark the skin at the location of the T1 spinous process. Participants were then re-positioned supine to permit MP by the examiners between the C1 and C7 levels. To do so, the examiners used the lateral aspect of their index fingers to apply over-pressure at end-range to the lateral aspect of the cervical articular pillars, using a mostly posterior to anterior and somewhat lateral to medial vector. This created extension, ipsilateral lateral flexion and contralateral rotation to the side of contact. This would be described as an end-feel method of MP[1,15], judging the quality of motion at end-range with pressure applied to one vertebra. The described over-pressure was equivalent to having taken each joint to end-range, left and right, as if to perform a traditional chiropractic adjustment commonly known as the modified rotary break, or more generically as the "supine proximal lateral index-transverse/articular pillar move."

After identifying the most hypomobile spinal site, the first examiner silently pointed out (by touching it) the location to a research assistant, who placed a small adhesive backed marker on the participant's skin at the indicated location. The examiner also whispered to the research assistant whether he was "very confident" or "not confident" in the finding of hypomobility. The research assistant then recorded the distance in centimeters from this marker to the mark on the T1 spinous process with a soft measuring tape, and also recorded the examiner's confidence rating. The marker was then removed and then the second and third examiners repeated the procedure, allowing approximately 2 minutes between observations.

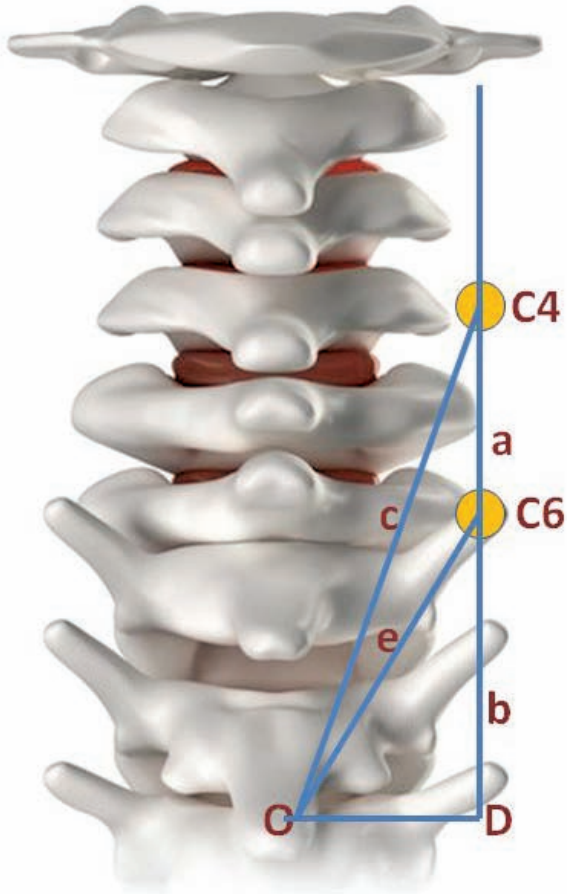

Figure 1. *Method for computing the distances between the examiner's locations for hypomobility*

Figure 1 illustrates our method for computing the distances between the examiner's locations for hypomobility, using findings at C4 and C6 as an example. Having obtained the distances from the hypomobile sites indentified by skin marks to the T1 spinous process, along lines O-C4

and O-C6, we used trigonometry to calculate distances a and b along the line D-C4. Line D-C4 is drawn 1.5 cm lateral to point O, based on a measurement taken from a dry spine. These calculations enabled us to transform the entire dataset of measurements taken from the spinous process of T1 to laterally situated hypomobile locations, to the actual vertical distances between examiners' findings. This heuristic calculation required three simplifying assumptions: (a) each level in the cervical spine was 1.8 cm in height[16] (even though vertebral height increases somewhat heading caudally); (b) the vertical length of the neck was uniform among participants; and (c) the skin marks were along a hypothetical line 1.5 cm from a hypothetical interspinous line.

We determined the spread of hypomobile findings across the range of C1-C7. We used the ICC statistic [2,1] (a two-way ANOVA model) to calculate group concordance and the root mean square error (RMSE) for all 27 analyzable participants, for all 3 examiners. We also calculated concordance among the 3 pair-wise combinations of examiners: 1&2 (n=29), 1&3 (n=27), and 2&3 (n=27). In each of the 4 datasets analyzed, the calculations were performed in three ways: for the sample as a whole (unstratified by doctor confidence), for a subset in which all examiners were confident in their findings, and for a subset in which at least one of the examiners lacked confidence. Thus, a total of 12 ICC and RMSE values were calculated. In addition, for each of these same 12 subsets we also calculated the mean absolute difference (MAD), the standard error of the mean (SEM), and 95% confidence intervals.

Results

Although 29 participants satisfied the inclusion criteria, two data points were not recorded for examiner 3 (one failure to record the distance from T1 to the hypomobile location, and one failure to record the confidence rating). As a result, some of the interexaminer reliability and MAD calculations (involving examiners 1&2) are based on n=29, and the others (any involving examiner 3) on n=27. Likewise, the calculations based upon the entire participant pool are based on n=27. Some mild erythema was often induced during examination, but was so dispersed that each examiner's identification of the most hypomobile location was judged to have been effectively masked from the other examiners.

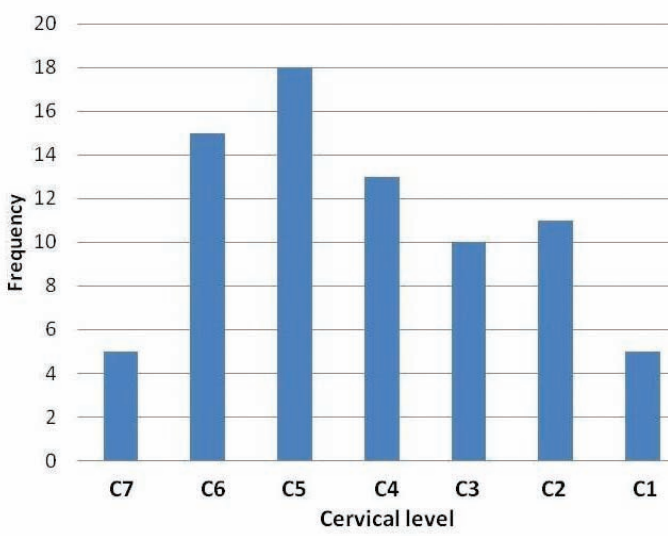

Figure 2. *Approximate distribution of hypomobility findings for all examiners and all participants*

Figure 2 is a histogram providing the approximate distribution of hypomobility findings for all examiners and all participants. Since the examiners did not attempt to nominate specific levels, but rather distances from a fixed point, we produced this histogram by distributing the actual measured locations into 7 bins within the C1-7 range. Since doing so required some simplifying assumptions, the histogram must be seen as a heuristic attempt to capture the results rather than representing exact numbers of findings at each segment. We assumed a uniform scale for the cervical spine with equally spaced segments; that at least one of the examiners' calls included C1; and that the vertical length of each participant's neck was uniform. Subject to these limits on interpretation, the hypmobile findings approximated a fairly smooth bell curve skewed to the left, with the peak frequency of hypomobile calls made at the the approximate level of C5.

The overall agreement on side of hypomobility was 44%. For examiners 1&2 kappa= –.39, p=.09; for examiners 1&3, kappa= –.28, p=.17; and for examiners 2&3, kappa= –.19, p=.39.

The percentage of participants for which at least one examiner lacked confidence was 28%, 17%, and 11% for examiners 1, 2, and 3 respectively.

Table 1 summarizes the data for subgroups 1-4 based on examiner comparisons: (1) all examiners combined;

Table 1
Interexaminer agreement

Set	Group description, sample size	Stratification state	Mean absolute difference	MAD 95% confidence interval		Standard error of mean	Intraclass correlation coefficient	
			MAD, cm	lower	upper	SEM	ICC	σ_e
1	All ex, n=27	non-stratified	1.35	1.12	1.57	0.12	0.61	1.22
	All ex, n=14	confident	1.13	0.88	1.39	.0.13	0.68	1.03
	All ex, 13	not confident	1.57	1.19	1.95	0.19	0.52	1.44
2	Ex 1&2, n=29	non-stratified	1.35	0.88	1.82	0.24	0.56	1.33
	Ex 1&2, n=16	confident	1.05	0.61	1.48	0.22	0.64	0.99
	Ex 1&2, n=13	not confident	1.72	0.85	2.59	0.44	0.50	1.70
3	Ex 1&3, n=27	non-stratified	1.28	0.96	1.60	0.17	0.68	1.10
	Ex 1&3, n=16	confident	1.47	1.06	1.88	0.21	0.57	1.19
	Ex 1&3, n=11	not confident	1.00	0.50	1.50	0.26	0.76	0.85
4	Ex 2&3, 27	non-stratified	1.38	1.82	1.73	0.18	0.60	1.19
	Ex 2&3, n=21	confident	1.23	0.81	1.61	0.20	0.67	1.10
	Ex 2&3, n=6	not confident	1.90	1.21	2.58	0.35	0.41	1.51

Ex=examiner; σ_e = root mean square error (RMSE), cm

(2) examiners 1&2; (3) examiners 1&3; and (4) examiners 3&4. In each subgroup, data are reported for 3 participant subsets: a non-stratified subset, a subset where all examiners were confident, and a subset where at least one examiner was not confident. The Shapiro–Wilk test was run on all subsets to confirm the populations were normally distributed prior to calculating reliability estimates.

Stratification by examiner confidence levels resulted in a series of relatively small subgroups for analysis. For the entire unstratified participant pool, ICC [2,1] = 0.61; the root mean square error (RMSE) indicated that accuracy of measurement was within 1.22 cm. The MAD in examiners' identification of the most hypomobile segment was 1.35 cm, 95% confidence interval: 1.12, 1.57 cm. Assuming an approximate 1.8 cm per cervical level[16], both the RMSE and MAD calculations suggest a mean interexaminer difference of about 75% of a vertebral level, clinically equivalent to having identified the same motion segment as being hypomobile.

Discussion

Possible explanations for the general poor reliability of previous motion palpation studies have included poor interexaminer spinal level localization leading to possible misreported discrepancies.[5,17] Some investigators avoided the spinal level numeration problem by having the examiners denote particular skin marks (applied beforehand by a research assistant) to represent the level of fixation, rather than attempting to number the vertebral levels felt to be fixated. Our study avoids the numeration problem by recording the hypomobile location as a distance from a landmark rather than as a spinal level in the usual sense of the term.

Our study allowed the examiners to determine the most hypomobile cervical location and rate their findings by degree of confidence. Using the typical scale for classifying ICC values (below 0.40 = poor, 0.40-0.59 = fair, 0.60-0.74 =good, above 0.75 = excellent)[14], interexaminer agreement for all participants, unstratified, was "good." Although the ICC values obtained in the present study were somewhat lower than the highest ICC reported in the aforementioned thoracic study[12], the results are generally comparable. MAD, the average of examiners' differences, was 1.35 cm, equivalent to having identified the same motion segment.

Our study was underpowered, resulting in subgroups that were exceedingly small. Thus, there was inadequate power to statistically determine significant differences in reliability among the confidence strata. Walter et al[18] developed a method for estimating the required sample size for ICC calculations, given the expected ICC, the lowest ICC that would be acceptable, and the number of raters. For example, using 3 examiners, had we been willing to accept ICC=0.4 ("fair"), and expecting ICC=.7 (comparable to a previous similar study[12], we would have needed a sample size of 20. For comparisons using 2 examiners, the required sample size would have been n=33.

Study designs that permit analyzing continuous group data rather than discrete paired observations may provide an increased ability to discern interexaminer agreement, as does (to some degree) allowing the examiners to rate their level of confidence in their findings. The methods used in most if not all previous MP studies asked the examiners to judge each spinal level as either moving normally or being hypomobile, and analyzed the data using the kappa statistic. Judging agreement by *how near* the identified locations are to one another, as we did, may be a more subtle and clinically relevant an assessing of agreement. It may better mirror how many (but certainly not all) manual therapists detect hypomobility in typical clinical settings: the palpator examines the spine for hypomobile sites that are relatively near the area of the patient's complaints.

Among the four dozen MP studies discussed in an annotated review of MP[2], Potter et al[19] were one of only two groups of investigators who used a most-hypomobile-segment paradigm similar to ours, and they also used ICC for the purposes of analysis. Since theirs was an intraexaminer study, unlike ours, and furthermore considered other examination findings in addition to MP to determine agreement, we cannot directly compare the results of their study with our own. Ghoukassian et al[20] also used a most fixated level protocol in an osteopathic study assessing the reliability of a percussive MP method developed by Johnston[21]. Even though they *could have* used ICC for the purposes of analysis, the investigators organized their data so as to use the kappa statistic, and doing so found negligible interexaminer agreement. When their data is reformatted so as to enable analysis using ICC, what becomes apparent is clinically relevant (although not high) agreement. Since the data in the Ghoukassian study[20] were presented such that they could be analyzed using either discrete or continuous statistical methods, in effect we have a direct head-to-head comparison whereby continuous analysis showed arguably more agreement than discrete analysis. An article re-analyzing this osteopathic study using continuous statistical methods is now in press.[22] It demonstrates that a head-to-head comparison of two measures of reliability, operating upon the same dataset, found greater reliability using continuous as compared with discrete analysis. However, this outcome may not be generalizable.

Interexaminer motion palpation studies generally assess agreement on the spinal level of hypomobility, but infrequently report the direction or side of restriction. Indeed, many by design constrain the examiners to study posterior to anterior glide only, or confine examination to one side only. In our study the overall agreement on the side of hypomobility was only 44%, although paradoxically, we found strong agreement on the spinal location of maximum hypomobility. Since the palpatory procedure involved applying mostly posterior to anterior pressure on the articular pillar and transverse process, the hard end-feel that the examiners perceived may have related more to vertebral resistance to extension (i.e., extension restriction) rather than axial rotation or lateral-to-medial translation. Under that assumption, the side contacted resulting in the judgement of extension hypomobility may have been perceptually unimportant.

We have no reason to think our results on side-specificity differ from those of other MP investigators, most of whom did not report their data on side-specificity. A cervical MP study by Cooperstein et al reported good examiner agreement on which participants in their study exhibited fixation, but not on the side.[23] There is some limited information available addressing cervical clinical outcomes as related to the side of intervention.[24-26]

Awareness seems to be growing that the kappa statistic has not been very useful in demonstrating examiner concordance in level-by-level study designs, and in fact needs to be made more useful by dint of using expanded definitions of agreement. For example, Abbot et al[27], after checking each lumbar level individually for instability, collapsed the data into just 2 levels, to which they then applied kappa statistics: "For analysis of clinical examination data, both clinical and radiographic data were then collapsed into two regions, corresponding to upper lum-

bar and lower lumbar. This was decided a priori, and considered necessary because there is considerable evidence that therapists are not sufficiently accurate in identifying specific segmental levels by palpation, although they are usually within one level (up or down) and are generally reliable at locating again a segment they had previously located." Heiderscheit et al[28] also realized that their prospects for finding segmental agreement using the kappa statistic was poor: "To account for potential segment level identification inaccuracies, an expanded definition of agreement was also used. Using this expanded definition, agreement with regard to the localization of findings was present if it was reproduced during the second examination session and located in the exact same spinal level or in a neighboring level (± 1 spinal segment)".

Our study design did not have to expand the bins to detect agreement. Examiners needed only to have been near each other in identifying the most hypomobile segment to be judged in agreement; and the closer they were, the higher the level of agreement. Our study was not designed to show that identifying the most fixated location is more clinically important than identifying a discrete level of care, nor address the clinical issues related to spinal manipulation based on any particular examination protocol. Rather, we attempted to show that examiners can be demonstrably concordant in identifying the most fixated location in the cervical spine, even though most prior cervical MP studies showed an unacceptable degree of agreement when assessing cervical motion level by level.

Limitations of the study

- Although we randomized colored slips of paper to determine the order of examiners for each participant, we did not record the examiner order. This precluded determining if there were order effects based on which examiner was first, second, or third.
- The research assistant may have erred to some degree in placing the marker at the spot indicated by the examiners and in performing the measurements. In addition, the research assistant was not blinded to the examiners' site designations.
- Since MP may alter the participant's joint movements (either increasing or decreasing end-range movement capability), using 3 examiners rather than the usual 2 might not have been a good design choice. It may have reduced the independence of the observations beyond what occurs using only 2 examiners.
- Lack of confidence in the examiners' rating of the most hypomobile motion segment might have come about in 2 different ways: an examiner might not have found *any* motion segment significantly hypomobile or an examiner may have found multiple segments significantly but indistinguishably hypomobile.
- The finding that the highest ICC recorded in this study occurred in a subgroup where at least one examiner was not confident was counter-intuitive and remains unexplained. That stated, examiner confidence levels appeared to have a modest impact in the study overall.
- Our study was underpowered, resulting in subgroups that were exceedingly small. Thus, there was inadequate power to statistically determine significant differences in reliability among the confidence strata. Small sample sizes[29] may increase the variability of examiners' observations.
- The study participants were largely asymptomatic, thus not reflective of symptomatic patients seeking care, jeopardizing the external validity in a manner that has been previously criticized[5,11]; lack of participant homogeneity[30] may increase variability. On the other hand there is some evidence that using more symptomatic participants does not appreciably change the outcome.[31]
- Without a reference standard we cannot confirm that there actually were any hypomobilities present in our study.

Conclusions

The palpators in this study of cervical end-feel MP exhibited good interexaminer agreement, with findings generally within one level of each other, despite having used industry-standard methods that previous studies had found mostly unreliable. Examiner confidence levels seemed to have a modest impact on the reliability of cervical spine

MP, but the study did not have enough power to address this clearly. There may be benefits to repeating the study on a sample of symptomatic patients.

With so many previous studies having been performed in educational institutions that used mostly asymptomatic and minimally-symptomatic participants, it is not surprising that the protocol of examining the spine segment by segment became established, and propagated in clinical situations even where more symptomatic participants were available. After all, minus a significant participant complaint, a level-by-level approach may have seemed more appropriate compared with the more targeted approach we suspect is used by field doctors, who might be expected to seek the most fixated or otherwise symptomatic segment lying within the field of primary complaint. Our clinical protocol of identifying the most hypomobile level may better represent the practice of some but not all clinicians using MP.

Investigators who perform research in educational institutions often use a convenience sample of minimally symptomatic students, often a relatively small sample, usually due to research infrastructural limitations. These investigators understand it would have been better to use a larger sample of more heterogeneous participants, and know in advance their study will never achieve a high score using rating instruments like QAREL[32] for reliability or QUADAS[33] for validity studies. Institutional investigators in such circumstances do need to make it clear that their studies have limited external validity, and readers should be cautious to not over-interpret their results.

The better reliability seen in our study compared with most previous motion palpation studies is not attributable to any improvements to the end-feel palpatory method, nor do they confirm a better method for identifying the most appropriate spinal site of care. We are not aware of any studies that report different outcomes for care based on examining every spinal level as compared with flagging the most relevant location within a patient's area of primary complaint. Therefore, these results do not call for clinicians to adopt new patient assessment methods or change their record-keeping protocols. They do suggest that researchers might consider designing their study protocols and research methods to explore reliability using the "most clinically relevant spinal site" protocol that some clinicians use. In fact, our results raise the possibility that the present inventory of mostly discrete (certainly for MP) reliability studies may underestimate clinically relevant examiner agreement, thereby unduly discouraging further research and clinician interest in such research. It may be possible to repeat many other interexaminer reliability studies, including studies of examination procedures other than MP (thermography, x-ray line marking, etc.) with similar design modifications that may more meaningfully assess examiner agreement than the mostly discrete analysis that has been used up until now.

We should not allow the confidence module of this study, given that it addressed an important clinical issue, to obscure our central finding. The interexaminer reliability for all 3 examiners, and for all participants, was good.

Acknowledgment

We would like to thank the Palmer Center for Chiropractic Research for its support.

References

1. Haneline MT, Cooperstein R, Young M, Birkeland K. Spinal motion palpation: a comparison of studies that assessed intersegmental end feel vs excursion. J Manip Physiol Ther. 2008 Oct;31(8):616-26.
2. Haneline M, Cooperstein R, Young M, Birkeland K. An annotated bibliography of spinal motion palpation reliability studies. J Can Chiropr Assoc. 2009 Mar;53(1):40-58.
3. Seffinger MA, Najm WI, Mishra SI, Adams A, Dickerson VM, Murphy LS, et al. Reliability of spinal palpation for diagnosis of back and neck pain: a systematic review of the literature. Spine. 2004 Oct 1;29(19):E413-25.
4. Stochkendahl MJ, Christensen HW, Hartvigsen J, Vach W, Haas M, Hestbaek L, et al. Manual examination of the spine: a systematic critical literature review of reproducibility. J Manip Physiol Ther. 2006 Jul-Aug;29(6):475-85, 85 e1-10.
5. Huijbregts PA. Spinal motion palpation: a review of reliability studies. J Man Manip Ther. 2002;10(1):24-39.
6. Haneline MT, Cooperstein R, Birkeland K. Spinal motion palpation: A comparison of studies that assessed intersegmental end-feel versus excursion. J Chiropr Educ. 2008;22(1):59-60.
7. Breen A. The reliability of palpation and other diagnostic methods. J Manip Physiol Ther. 1992;15(1):54-6.
8. Dishman RW. Static and dynamic components of the chiropractic subluxation complex: a literature review [see comments]. J Manip Physiol Ther. 1988;11(2):98-107.
9. Haas M, Panzer D, Raphael R. Reliability of manual end-play palpation of the thoracic spine. Chiropr Tech. 1995;7(4):120-4.

10. Troyanovich SJ, Harrison DD. Motion Palpation: It's time to accept the evidence. J Manip Physiol Ther. 1998;21(8):568-71.
11. Hestbaek L, Leboeuf-Yde C. Are chiropractic tests for the lumbo-pelvic spine reliable and valid? A systematic critical literature review. J Manip Physiol Ther. 2000;23(4):258-75.
12. Cooperstein R, Haneline M, Young M. Interexaminer reliability of thoracic motion palpation using confidence ratings and continuous analysis. J Chiropr Med. 2010 Sep;9(3):99-106.
13. Haas M. Statistical methodology for reliability studies. J Manip Physiol Ther. 1991 Feb;14(2):119-32.
14. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005 Mar;85(3):257-68.
15. Brown J, Cooperstein R. Why motion palpation is so confounding. J Am Chiropr Assoc. 2001;38(10):34-6.
16. Gilad I, Nissan M. Sagittal evaluation of elemental geometrical dimensions of human vertebrae. J Anat. 1985 Dec;143:115-20.
17. Billis EV, Foster NE, Wright CC. Reproducibility and repeatability: errors of three groups of physiotherapists in locating spinal levels by palpation. Man Ther. 2003 Nov;8(4):223-32.
18. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998 Jan 15;17(1):101-10.
19. Potter NA, Rothstein JM. Intertester reliability for selected clinical tests of the sacroiliac joint. Phys Ther. 1985;65(11):1671-5.
20. Ghoukassian M, Nicholls B, McLaughlin P. Inter-examiner reliability of the Johnson [sic] and Friedman percussion scan of the thoracis spine. J Am Osteopath Assoc. 2001;4(1):15-20.
21. Johnston WL, Allan BR, Hendra JL, Neff DR, Rosen ME, Sills LD, et al. Interexaminer study of palpation in detecting location of spinal segmental dysfunction. J Am Osteopath Assoc. 1983 Jul;82(11):839-45.
22. Cooperstein R. Interexaminer reliability of the Johnston and Friedman percussion scan of the thoracic spine: secondary data analysis using modified methods. J Chiropr Med. 2012;11(3):154-159.
23. Cooperstein R, Gardner R, Nansel D. Concordance of two methods of motion palpation with goniometrically-assessed cervical lateral flexion asymmetry. International Conference on Spinal Manipulation; 1991; Arlington, VA.: FCER.
24. van Schalkwyk R, Parkin-Smith GF. A clinical trial investigating the possible effect of the supine cervical rotatory manipulation and the supine lateral break manipulation in the treatment of mechanical neck pain: a pilot study. J Manip Physiol Ther. 2000 Jun;23(5):324-31.
25. Cilliers KI, Penter CS. Relative effectiveness of two different approaches to adjust a fixated segment in the treatment of facet syndrome in the cervical spine. J Neuromusculoskeletal System. 1998;6(1):1-5.
26. Hubka MJ, Phelan SP, Delaney PM, Robertson VL. Rotary manipulation for cervical radiculopathy: observations on the importance of the direction of the thrust. J Manip Physiol Ther. 1997;20(9):622-7.
27. Abbott JH, McCane B, Herbison P, Moginie G, Chapple C, Hogarty T. Lumbar segmental instability: a criterion-related validity study of manual therapy assessment. BMC Musculoskelet Disord. 2005;6:56.
28. Heiderscheit B, Boissonnault W. Reliability of joint mobility and pain assessment of the thoracic spine and rib cage in asymptomatic individuals. J Man Manip Ther. 2008;16(4):210-6.
29. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med. 1990;20(5):337-40.
30. Ekeberg OM, Bautz-Holter E, Tveita EK, Keller A, Juel NG, Brox JI. Agreement, reliability and validity in 3 shoulder questionnaires in patients with rotator cuff disease. BMC Musculoskelet Disord. 2008;9:68.
31. DeCamp Jr. N, editor. Objective analysis of the lumbo-sacral complex and occiput. Sacro Occipital Research Society International Symposium; 1987; San Diego, California.
32. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol. 2010 Jan 5;63(8).
33. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003 Nov 10;3:25.