Journal of Chiropractic Medicine (2012) **11**, 154–159

ELSEVIER

JCM Journal of Chiropractic Medicine

www.journalchiromed.com

Interexaminer reliability of the Johnston and Friedman percussion scan of the thoracic spine: secondary data analysis using modified methods

Robert Cooperstein MA, DC*

Professor, Coordinator of Research and Technique, Palmer Center for Chiropractic Research, Palmer Chiropractic College West, San Jose, CA

Received 21 January 2012; received in revised form 31 May 2012; accepted 5 June 2012

Key indexing terms:
Palpation;
Percussion;
Reproducibility of results;
Biomechanics;
Range of motion;
Articular;
Chiropractic

Abstract

Objective: The purpose of this study is to perform a secondary analysis using modified methods of previously reported data to analyze the amount of examiner concordance in the Johnston and Friedman percussion scan of the most fixated spinal level.

Method: A 2001 study evaluated interexaminer reliability of the percussive method of Johnston and Friedman for detecting altered segmental mobility (somatic dysfunction, spinal/segmental dysfunction, or chiropractic subluxation) in the thoracic spine. The original reported level of agreement using the κ statistic for discrete measures was only 0.07, judged "slight." The data were reformatted to permit recalculating the degree of interexaminer agreement using the intraclass correlation coefficient statistic, which uses continuous analysis, unlike κ that performs discrete analysis. Following an initial calculation, the data were modified to reflect the caudally increasing vertebral height of the thoracic vertebrae.

Results: The reformatted and modified data, intraclass correlation coefficient (2,1) = 0.253 (0.100,0.482), showed the findings as "poor," which is better interexaminer agreement for percussion motion palpation than the original reported κ value judged as "slight."

Conclusions: Reanalyzing the data using an alternative statistical method showed greater interexaminer reliability than was originally reported. This secondary analysis demonstrates how study results may vary depending on the experimental design and statistical methods chosen for analysis.

Introduction

The study by Ghoukassian et al[1] evaluated interexaminer reliability of the Johnston and Friedman

* Palmer Chiropractic College West, 90 East Tasman Dr, San Jose, CA 95134. Tel.: +1 408 944 6009; fax: +1 408 944 6118.
E-mail address: Cooperstein_r@palmer.edu.

1556-3707/$ – see front matter © 2012 National University of Health Sciences.
http://dx.doi.org/10.1016/j.jcm.2012.06.001

percussive scan of the thoracic spine.[2] This osteopathic palpatory procedure purports to detect increased tissue tension and joint hypomobility, both considered elements of somatic dysfunction, spinal/segmental dysfunction, or chiropractic subluxation. The procedure consists of the thumb and third finger of the examiner straddling the thoracic spinous processes so as to contact the paravertebral muscles, and then proceeding down the vertebral column deploying one percussive strike per segment. Vertebral movement at each level is assessed and compared with movement at segments above and below. Ghoukassian et al describe the findings being "motion restriction results in increased tension and decreased elasticity of the segmental musculature, leading to a decreased rebound to the percussion stroke."[1] It is unknown how many doctors of chiropractic may use a similar manual percussive treatment method, although one similar example may be the Pro-Adjuster, a computer-controlled mechanical percussive instrument some use to identify and treat fixated segments.[3]

For this study, I wanted to explore a different question than the one addressed by the original investigators. Although the original study evaluated the probability of exact agreement, for this reanalysis, the process will analyze *how far apart* the examiners had been in their determinations of the most fixated spinal level. Review of the data suggested that they could be reorganized so as to determine the proximity of the examiners' findings. Therefore, the purpose of this study is to perform a secondary analysis of previously reported data to analyze the amount of examiner concordance in palpation of the most fixated spinal level.

Methods

The study of Ghoukassian et al recruited 19 asymptomatic male volunteers (mean age, 22 years) and 10 senior postgraduate osteopathic students as examiners, each having had at least 2 years' experience using the percussive method. The examiners had 2 training sessions to standardize the protocol. Each examiner then examined each participant, identifying the "most significant area of altered tissue tension,"[1] the level that manifested the least rebound to the percussive stroke between T1 and T12. Doctors of chiropractic may use similar terms such as *loss of springiness*, *decreased intersegmental motion*, or *fixation* to convey the same clinical impression.[4]

Using the κ statistic, the investigators reported interexaminer agreement to be 0.07 ($P < .01$), which would be judged only "slight."[5] Ghoukassian et al concluded: "This result suggests that the inter-examiner reliability of this examination procedure remains questionable when used alone."[1] These results seemed far less impressive than those reported in the predecessor study of Johnston and Friedman,[2] who had reported 79% interexaminer agreement. However, simply reporting percentage agreement among examiners does not correct for chance agreement and thus may overstate the level of agreement.[6] Using κ to calculate interexaminer reliability is more interpretable because it corrects for chance agreement.[7]

The data from Ghoukassian et al are reported in Table 1, adapted from the published article. Each cell reports the number of examiners out of 10 who found a given level to be the most fixated for each of the 19 participants. Although the κ value was very low, suggesting low reliability, simple inspection of Table 1 tells a somewhat different story. For example, in the case of participant 10, 6 of 10 examiners found T4 to be the most fixated segment; and for participant 19, all of the examiners found the most fixated segment to lie between T4 and T8. Although *exact* examiner agreement was generally speaking infrequent, there was noticeable agreement on the *approximate location* of somatic dysfunction.

Table 1 Original data, number of examiners finding given vertebral level the most fixated

	Vertebral level											
Participants	1	2	3	4	5	6	7	8	9	10	11	12
1			1	3	1	1	2	1		1		
2		2		3					2	2	1	
3			1					2		3	4	
4			1		2	2	1			1	3	
5		1	3	1	1	1	1	1	1			
6	1	5	1	1			1		1			
7	2	1	2	2	1	1			1			
8	1			1		1	3	2	2			
9				1	3	1		4	1			
10		1		6	1			1		1		
11					3	3	2	1	1			
12	2	3	1	2			1					1
13			3	1	2		2				2	
14					2	2			2	3		1
15						3	3	3		1		
16		1			2		1			2	2	2
17		3	3	1				1			2	
18	1			1		2	1	1	2		2	
19				2	3	1	3	1				

Table 2 Reformatted data, segments distant from C7, examiner order random

	Vertebral level									
Participants	1	2	3	4	5	6	7	8	9	10
1	3	4	4	4	5	6	7	7	8	10
2	2	2	4	4	4	9	9	10	10	11
3	3	8	8	10	10	10	11	11	11	11
4	3	5	5	6	6	7	10	11	11	11
5	2	3	3	3	4	5	6	7	8	9
6	1	2	2	2	2	2	3	4	7	9
7	1	1	2	3	3	4	4	5	6	9
8	1	4	6	7	7	7	8	8	9	9
9	4	5	5	5	6	8	8	8	8	9
10	2	4	4	4	4	4	4	5	8	10
11	5	5	5	6	6	6	7	7	8	9
12	1	1	2	2	2	3	4	4	7	12
13	3	3	3	4	5	5	7	7	11	11
14	5	5	6	6	9	9	10	10	10	12
15	6	6	6	7	7	7	8	8	8	10
16	2	5	5	7	10	10	11	11	12	12
17	2	2	2	3	3	3	4	8	11	11
18	1	4	6	6	7	8	9	9	11	11
19	4	4	5	5	5	6	7	7	7	8

Reformatting the data was done so that examiner agreement could be analyzed using the intraclass correlation coefficient (ICC) regarding the thoracic spinal levels to comprise an estimated interval scale ("estimated" because the intervertebral distances were not equal, increasing caudally). Table 2 was derived from the data in Table 1 by using C7 as an arbitrary reference point for calculating the relative location of the most fixated segment as determined by each of the multiple examiners.

For the purpose of analysis, it was necessary to measure how close the examiners were for each participant. Rather than directly calculating these distances by using the equivalent of a ruler, it was more convenient to get the data into a statistics program by calculating the location of the fixations from an arbitrary point. For example, if C3 and C4 were found fixated by examiners 1 and 2, respectively, then their ratings might have been directly calculated to be 2 cm apart using the arbitrary metric that 1 vertebral level = 2 cm. Alternatively, C3 could be measured to be 6 cm from C7, and C4 to be 8 cm from C7. Subtracting, we would indirectly derive the same distance between the 2 examiners' fixation locations: 2 cm.

Turning to the data in Table 1, it can be seen that, in the case of participant 1, no examiner found T1 or T2 to be the most fixated level, but one examiner did find T3 the most fixated. Because T3 is 3 levels distant from C7, the number "3" is entered into the first cell in the table, representing the place the first examiner found the first participant to be the most fixated. Proceeding in a similar fashion, Table 1 tells us that 3 examiners found T4, which is 4 levels from C7, the most fixated segment. Thus, the number "4" is entered into the next 3 cells of the first row. By extending this counting procedure for all participants, the author was able to derive Table 2 from the original data in Table 1. The

Table 3 Corrected data, computed approximate segments distant from C7, examiner order random

	Vertebral level									
Participants	1	2	3	4	5	6	7	8	9	10
1	3.28	4.55	4.55	4.55	5.92	7.38	8.93	8.93	10.58	14.14
2	2.09	2.09	4.55	4.55	4.55	12.31	12.31	14.14	14.14	16.06
3	3.28	10.58	10.58	14.14	14.14	14.14	16.06	16.06	16.06	16.06
4	3.28	5.92	5.92	7.38	7.38	8.93	14.14	16.06	16.06	16.06
5	2.09	3.28	3.28	3.28	4.55	5.92	7.38	8.93	10.58	12.31
6	1.00	2.09	2.09	2.09	2.09	2.09	3.28	4.55	8.93	12.31
7	1.00	1.00	2.09	3.28	3.28	4.55	4.55	5.92	7.38	12.31
8	1.00	4.55	7.38	8.93	8.93	8.93	10.58	10.58	12.31	12.31
9	4.55	5.92	5.92	5.92	7.38	10.58	10.58	10.58	10.58	12.31
10	2.09	4.55	4.55	4.55	4.55	4.55	4.55	5.92	10.58	14.14
11	5.92	5.92	5.92	7.38	7.38	7.38	8.93	8.93	10.58	12.31
12	1.00	1.00	2.09	2.09	2.09	3.28	4.55	4.55	8.93	18.07
13	3.28	3.28	3.28	4.55	5.92	5.92	8.93	8.93	16.06	16.06
14	5.92	5.92	7.38	7.38	12.31	12.31	14.14	14.14	14.14	18.07
15	7.38	7.38	7.38	8.93	8.93	8.93	10.58	10.58	10.58	14.14
16	2.09	5.92	5.92	8.93	14.14	14.14	16.06	16.06	18.07	18.07
17	2.09	2.09	2.09	3.28	3.28	3.28	4.55	10.58	16.06	16.06
18	1.00	4.55	7.38	7.38	8.93	10.58	12.31	12.31	16.06	16.06
19	4.55	4.55	5.92	5.92	5.92	7.38	8.93	8.93	8.93	10.58

numbers entered into cell of Table 2 tell us how far each of the 10 examiners was from C7 for each of the 19 participants. We cannot determine which examiner identified any particular segment as the most fixated; but we do not need such information to calculate the ICC, which, unlike some other correlation measures, operates on data that are structured as groups. By comparison, the κ statistic operates on data that are structured as paired observations.

Having calculated an initial ICC for the derived data in Table 2, the author then modified those data to reflect the fact that the thoracic spinal levels do not comprise a cardinal scale per se because each successive thoracic level headed caudally is more than one incremental unit distant from C7. Taking an approximate measurement from a dry spine, the height of T1 was very close to two-thirds the height of T12. This measurement allowed transforming Table 2 into Table 3, which takes into account the increasing height of the caudal thoracic vertebral levels. This in turn allowed calculating a modified, revised ICC for the study of Ghoukassian et al.[1]

Results

Applied to the transformed data in Table 2, uncorrected for increasing caudal height of thoracic vertebrae, ICC (2,1) = 0.262 (0.104, 0.494). This would be judged "poor" according to the following scale: above 0.75 = good reliability, 0.40 to 0.75 = fair to good reliability, below 0.40 = poor reliability.[8] Applied to the transformed data in Table 3, corrected for the increasing caudal height of thoracic vertebrae, the revised ICC value was ICC (2,1) = 0.253 (0.100,0.482), also judged "poor."

Discussion

Although statistical beauty is always in the eyes of the beholder, the author believes the "poor" agreement attained in this reanalysis using continuous measures and ICC value is more impressive than the "slight" agreement previously reported using the κ statistic. One can only wonder what the results of other spine-related interexaminer reliability studies would have been had they used a continuous measures methodology, amenable to analysis with ICC. Given the generally low levels of agreement that in fact have been reported in reliability studies in the manual arts, the way forward may lie less in improving examination techniques and more in improving the methods by which we assess examiner concordance. One can only wonder what might have been the results had a continuous measures methodology been used in virtually countless studies of radiograph marking, leg checking procedures, thermographic evaluation, motion palpation (MP), and other low-tech patient evaluation methods.

In most palpation studies, the examiners decide for each segment whether they find it fixed or not (springy or not, having a hard end-feel or not, etc). Such studies are typically evaluated using the κ statistic. For example, Haas et al[9] performed a thoracic end-feel palpation study that reported $\kappa = 0.14$. In their thoracic study, Ghoukassian et al[1] also used κ even though the data were amenable to analysis with ICC.

In an interexaminer reliability study, the best that examiners can do is to exactly agree in all their ratings. The next best outcome would be that they almost agree in most of their ratings; that would indicate more reliability than purely and simply disagreeing most of the time. The ICC statistic is designed to calculate how similar examiners' ratings are, whereas the κ statistic is designed to calculate to what degree there is exact agreement. The purpose of the present study was to illustrate, using the vehicle of a study on percussive palpation, that ICC may more clearly mirror interexaminer reliability by setting a more clinically realistic and relevant level for defining "agreement." Because some fixations may result from contractures of muscles and ligaments that span several segments, it may be more relevant to define agreement as having close proximity rather than strict segmental concordance. In principle, the use of the κ statistic can be made more liberal by considering examiners in agreement when their calls are within ± 1 spinal segment of each other. For example, Christensen et al,[10] using such a definition of agreement, reported intraexaminer reliability to be good (κ = 0.59 to 0.77), whereas interexaminer reliability was low (κ = 0.24 and 0.22).

The clinical consequence of examiners "almost agreeing" as compared with exactly agreeing is no doubt variable. One would prefer exact agreement if examiners were making the call on the presence of a life-threatening illness, but *almost agreeing* may be acceptable when deciding what spinal level might be best adjusted in a patient who has uncomplicated musculoskeletal pain.

The ICC levels computed for original and modified data in Tables 2 and 3 were 0.262 and 0.253, which both must be judged to be "poor." The ICC values in this range, although they do not strongly support the present utility of the examination procedure, are high

enough to warrant further research toward modifying the method and improving upon the results. The results as reported by Ghoukassian et al in the original study, $\kappa = 0.07$,[1] was not construed to warrant further attention. The present reanalysis and reinterpretation of the original data serve as a reminder that our judgment on the merits of a patient assessment method could be rather sensitive to the mode of analysis chosen to interpret study outcomes.

In continuous analysis, the measuring device has units—1 in, 1 mm, perhaps 1 vertebral level, as the case may be. In the first iteration, for the sake of simplicity, the author assumed that, in effect, the original investigators had used a measuring device with a resolution of 1 vertebral level. In the second iteration, the author took into account that the thoracic levels are not spaced at equal increments, and corrected the data for measurement bias. This did not appreciably change the results.

In our own study on the interexaminer reliability of thoracic MP,[11] we deployed a methodology very similar to that used by Ghoukassian et al.[1] Whereas they used an upright percussive examination technique, we used a prone end-feel method to evaluate posterior to anterior glide in the thoracic spine. In our study, each of the 2 examiners measured the distance between the most fixated level between T3 and T11 to an arbitrary fixed point on the sacral base. Because we directly recorded our findings in centimeters rather than as discrete vertebral levels, we did not have to convert spinal levels to centimeters to compute ICCs.

Our study included another very important design feature: we allowed the examiners to rate their level of confidence in their findings as either "very confident" or "not confident." By doing so, we were able to analyze a subset of the total sample in which both doctors were confident in their findings: ICC = 0.827 (95% confidence interval, 0.626-0.925). This was judged to be "good to excellent" reliability; we are not aware of any other study that reported better interexaminer reliability, at least among studies where there was no dialogue allowed between the study participants and the examiners.

For all participants combined in our study, unstratified by doctors' confidence ratings, ICC = 0.311 (95% confidence interval, 0.046-0.536). This lower ICC is rather close to the ICC achieved in the study of Ghoukassian et al.[1] Therefore, we may hypothesize that these latter investigators, had they allowed stratification of the doctors' calls as we did, may have achieved a higher level of interexaminer agreement in a hypothetical confident subgroup.

The prior study is one of the few that achieved high reliability and used palpation only; there was no verbal interaction with the participants, as occurred in a well-known cervical MP study[12] that also showed high reliability. Our study made no breakthrough in the palpatory method. Rather, we used a method of calculating interexaminer agreement. It was this experience that suggested that the results in the study of Ghoukassian et al[1] were possibly due to choice of study design and method of data analysis.

Among other MP studies, only 2[13,14] used a most fixated segment paradigm similar to that of Ghoukassian et al.[1] Although Potter and Rothstein, like ourselves, used ICC to assess concordance, theirs was an intraexaminer study. They used findings other than MP to assess agreement. For these reasons, we cannot compare their study's findings with our own or with those of Ghoukassian et al.

Limitations of the study

As a secondary analysis of previously published data, this present study could not control for the limitations of the parent study.[1] Data obtained from the parent study were extracted from the published article and were not from original raw data spreadsheets; thus, there may have been errors introduced by using these methods. A limitation of this secondary analysis was the variable distance between vertebral levels; although the author did introduce a heuristic correction to reduce the errors, this did not have much impact.

Suggestions for future studies

At present, the clinical value of MP in identifying optimal locations to target manipulative or other therapeutic procedures is unknown. To our knowledge, it has not been demonstrated that the information provided by MP improves the outcome of clinical care. On the contrary, one randomized clinical trial (although it itself had some limitations) suggested that the information provided by MP did not immediately improve the outcome of care.[15] Thus, because of lack of information, it is unknown if findings from spinal palpation contribute to treatment procedures and clinical outcomes.

In regard to future studies for interexaminer reliability, altering the experimental design, in particular so as to permit continuous analysis, may improve the ability to detect examiner agreement. Using a continuous measures methodology for gathering and analyzing data could improve the results of other interexaminer reliability studies in the manual arts. As

well, future studies should include subgrouping by examiner confidence levels.

Ultimately, more reliable methods of patient evaluation are needed. Underdevelopment of good assessment methods, both reliable and valid, will ultimately prevent us from properly selecting participants likely to benefit from chiropractic care in outcome studies. There is an old adage that there are 3 factors in obtaining a good treatment outcome: patient selection, patient selection, and patient selection. Outcome studies in the manual arts, not to mention the derivation of clinical prediction rules[16] studies, are likely to continue underperforming unless and until we adopt more realistic and clinically relevant means of operationally defining agreement.

Conclusion

Reanalyzing the data using an alternative statistical method showed greater interexaminer reliability than was originally reported. This secondary analysis demonstrates how study results may vary depending on the experimental design and statistical methods chosen for analysis. Intraclass correlation coefficient may more clearly mirror interexaminer reliability by setting a more clinically realistic and relevant level for defining "agreement."

Funding sources and potential conflicts of interest

No funding sources or conflicts of interest were reported for this study.

References

1. Ghoukassian M, Nicholls B, McLaughlin P. Inter-examiner reliability of the Johnson [sic] and Friedman percussion scan of the thoracic spine. J Am Osteopath Assoc 2001;4(1):15-20.
2. 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;82(11):839-45.
3. Zhang J, Haselden P, Tepe R. A case series of reduced urinary incontinence in elderly patients following chiropractic manipulation. J Chiropr Med 2006;5(3):88-91.
4. Haneline MT, Cooperstein R, Young M, Birkeland K. Spinal motion palpation: a comparison of studies that assessed intersegmental end feel vs excursion. J Manipulative Physiol Ther 2008;31(8):616-26.
5. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):360-3.
6. Haas M. Statistical methodology for reliability studies. J Manipulative Physiol Ther 1991;14(2):119-32.
7. Haneline M, Cooperstein R. Weighing the reliability and validity of clinical tests. J Am Chiro Assoc 2006;43(7): 19-22.
8. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 2000.
9]. Haas M, Panzer D, Raphael R. Reliability of manual endplay palpation of the thoracic spine. Chiropr Tech 1995;7(4): 120-4.
10. Christensen HW, Vach W, Vach K, Manniche C, Haghfelt T, Hartvigsen L, et al. Palpation of the upper thoracic spine: an observer reliability study. J Manipulative Physiol Ther 2002;25(5):285-92.
11. Cooperstein R, Haneline M, Young M. Interexaminer reliability of thoracic motion palpation using confidence ratings and continuous analysis. J Chiropr Med 2010;9(3): 99-106.
12. Jull G, Bogduk N, Marsland A. The accuracy of manual diagnosis for cervical zygapophysial joint pain syndromes. Med J Aust 1988;148(5):233-6.
13. Haneline M, Cooperstein R, Young M, Birkeland K. An annotated bibliography of spinal motion palpation reliability studies. JCCA J Can Chiropr Assoc 2009;53(1):40-58.
14]. Potter NA, Rothstein JM. Intertester reliability for selected clinical tests of the sacroiliac joint. Phys Ther 1985;65(11): 1671-5.
15. Haas M, Groupp E, Panzer D, Partna L, Lumsden S, Aickin M. Efficacy of cervical endplay assessment as an indicator for spinal manipulation. Spine 2003;28(11):1091-6 [discussion 6].
16. Cooperstein R. The derivation and validation of clinical prediction rules: the quest to reduce clinical uncertainty. J Am Chiro Assoc 2011;48(4):8-13.