

Palpatory accuracy: time to reflect

Numerous studies indicate a relatively poor degree of accuracy in standard assessment and palpation tests. It might be imagined that this trend would be cause for concern amongst responsible clinicians and educators, and yet apart from the regular appearance of ever more studies questioning the value of individual manual assessment procedures, the methods in question continue to be widely used and taught as the basis for the development of subsequent treatment protocols.

Lord and Bogduk (1996) have stated 'There have been many claims regarding the accuracy of manual diagnosis but few data. Only one study (Jull & Bogduk 1988) compared manual diagnosis to the criterion standard of local anaesthetic blocks. The authors found the sensitivity and specificity of the manual examination technique to be 100%. The manual therapist correctly identified all patients with proven joint pain, the symptomatic and asymptomatic segments. The ability of other manual examiners to replicate these results has not been tested'. When research is carried out relative to intra- and inter-rater reliability (see examples below relating to

individual practitioner's degree of consistency in performance of assessment methods [intra], as well as agreement of findings between different practitioners performing the same assessment [inter], the results are commonly poor, and often bad. This sort of documentation might create a defensive response amongst many who claim to rely on their palpation skills for accurate assessment of dysfunction, however the evidence is mounting that much clinical assessment and palpation is anything but reliable – especially when individual test procedures are relied on to decide on treatment strategies.

Before listing a selection of some such evidence it is worth reflecting that other forms of assessment, such as radiography, can also be held up as examples of relative inaccuracy.

For example, Christensen et al. (2001) conducted a study to evaluate the accuracy of radiographic interpretation of a posterolateral spinal fusion mass. They note that, in general, the literature describing the classification criteria used for radiographic interpretation of spinal posterolateral fusion has serious deficiencies. The study they designed was a prospective randomized clinical study in which four experienced observers evaluated radiographs of posterolateral fusion masses. The mean interobserver agreement was 86% (Kappa 0.53), and the mean intraobserver

agreement was 93% (Kappa 0.78).* No difference in interobserver and intraobserver agreement was found between patients with and without supplementary pedicle screw fixation. All mean Kappa values were classified as fair or good. And yet the conclusion was that it is 'extremely difficult' to interpret radiographic images of lumbar posterolateral fusion.

Dr. Zachary Comeaux (see article in this issue on page 245) comments on this study that 'the interpretation of X-rays... in the Christensen article, demonstrated actual Kappa scores of 0.53 and 0.78, which is in the same range as in some of the palpation studies which are questioned as to accuracy'. Professor Comeaux DO further observes, 'Any clinician who reads the radiologists report of imaging procedures recognizes the interpretation challenge and variation of interpretation with the same films. Why should palpation be expected to be any less of an art?' (personal communication June 2001).

Examples of palpation studies which reflect poorly on accuracy include:

1. Standing flexion test for sacroiliac dysfunction: A study by Vincent-Smith and Gibbons (1999) showed that, 'the

*Kappa greater than 0.75 = excellent. Kappa between 0.4 and 0.75 = good reproducibility. Kappa less than 0.4 = poor reproducibility (Fleiss 1981).

- reliability of the standing flexion test as an indicator of SIJ dysfunction still remains questionable'. Inter-examiner reliability showed 'a mean percentage agreement of 42% and a K coefficient of 0.052 demonstrating statistically insignificant reliability'. Intra-examiner reliability data demonstrated a mean percentage agreement of 68% and a K coefficient of 0.46 indicating moderate reliability.
2. Slipman et al. (1998) investigated the predictive value of SI joint 'provocation' tests, as compared with what they describe as the medical 'gold standard' approach, of a joint block injection. Fifty patients were selected to be tested by joint block if they tested positive using at least three manual methods. The manual provocation tests always included Patrick's F-AB-ER-E test, as well as direct palpation for pain in the ipsilateral sacral sulcus, plus one other from the wide range of choices available, such as pain provocation by means of the transverse anterior distraction compression test, or transverse posterior distraction test, or Gaenslen's test. The working hypothesis was that if the joint block injections (performed using fluoroscopically guided needling) effectively eliminated SI pain, the manual assessment had been accurate. The results showed that 30 of the 50 patients were relieved of symptoms by 80% or more by means of joint block, whereas 20 achieved less than 80% relief. A 60% degree of accuracy in identifying SI joint syndrome was therefore noted using manual testing, in this study. Conclusion: 'Our results do not support the use of provocative SIJ manoeuvres to confirm SIJ syndrome diagnosis'.
 3. Static palpation was performed for the location of three standard landmarks used in assessment of the SI joint, posterior superior iliac spine (PSIS), sacral sulcus (SS), sacral inferior lateral angle (SILA) on 10 asymptomatic individuals by 10 senior osteopathic students. Intra-examiner agreement results were: for SILA less than chance; and for PSIS and SS slight to moderate. Inter-examiner results were 'slight'. Conclusion: 'The poor reliability of clinical tests involving palpation may be partially explained by error in landmark location' (O'Haire & Gibbons 2000).
 4. For inter-therapist reliability, 14 physiotherapists were tested to see whether they could identify accurately specific spinal levels (which had been pre-marked with an ultra violet ink) by palpation of five normal subjects, while for intra-therapist reliability three therapists palpated five normal subjects. There was only fair agreement between physiotherapists (K = 0.28); however, intra-therapist agreement ranged from substantial to almost perfect (K = 0.61, 0.70, 0.90). 'The lack of inter-therapist reliability suggests that further research into reliability of spinal palpation is required if it is to remain an important component of spinal therapy' (McKenzie & Taylor 1997).
 5. Gillet's test, which assesses sacroiliac mobility, is widely used, albeit with variations in manual contacts (including L5 spinous process, different aspects of the PSIS, sacral spinous processes etc) as the patient removes weight from, and raises one leg or the other. A study was conducted in which some 40 symptomatic and asymptomatic subjects were evaluated using variations on the test. Their conclusion was that 'The Gillet test – as performed in this study – does not appear to be reliable' (Meijne et al. 1999).
 6. A complex study involved chiropractic assessment of possible lumbopelvic dysfunction, low back pain (LBP), in 83 pairs of twins. The study evaluated use of a variety of different tests each of which has the option of a 'yes', a 'no' or a 'don't know' answer to the question of 'is there dysfunction?' (Observational: Stork test, non-reversal of lordosis; antalgic posture, gearbox movement on flexion; pain-on-movement tests involving flexion, extension, left and right side bending and rotation; pain provoking tests: forced extension, percussion of spinous processes). The tests were evaluated to see if they met the following criteria: (1) logical pattern in relation to LBP status; (2) high sensitivity; (3) high specificity; (4) high positive; and (5) high negative predictive values. The researchers results were: 'None of the single tests fulfilled all five of the criteria. However, the variable of at least one painful movement did fulfil all of the criteria. Each of the individual lumbar movements (except left rotation) fulfilled all the criteria but one (high sensitivity). The gearbox test was only present in one person, who was correctly classified as 'LBP today'. Conclusion: 'although no individual test was accurate, the diagnostic discrimination on the basis of these tests was satisfactory' (Lebouef-Yde & Kyvik 2000).

A need to compound evidence

This last reported study suggests that clinically, no single test should

be used diagnostically, but that a picture should be built based on compounding of evidence. Accuracy may then be sufficient for the purpose of devising subsequent treatment protocols.

Lee (1999) agrees with Bogduk (1997) that ideally a 'biomechanical diagnosis require biomechanical criteria' and that 'pain on movement is not that criteria'.

The fact that there is relatively poor inter-tester reliability when applying tests does not necessarily negate the value of these tests, *merely the efficiency with which they are applied*, according to Lee who continues, 'The tests for spinal and sacroiliac function (i.e., mobility/stability, not pain) continue to be developed and hopefully will be able to withstand the scrutiny and rigor of scientific

research and take their place in a clinical evaluation which follows a biomechanical and not a pain model.' Discussing the value of tests, she notes that while individually, in isolation, some may fail evaluation as to their reliability and validity, when combined into a sequence of numerous evaluation strategies, and especially when '*a clinical reasoning process is applied to their findings*', they combine to offer a logical biomechanical diagnosis, and 'without apology, they continue to be defended.'

JBMT intends to publish a forum for debate and discussion of these issues, along the lines of the extremely useful 'Muscle debate' (JBMT 4(4): 225-241), which carried the perspectives of world-renowned experts on the topic of muscle

classification issues.

A series of questions will be posed to selected individuals in the various manual therapy (bodywork) professions, including chiropractic, manual medicine, massage therapy, osteopathy, physiotherapy and Rolting. Short Succinct answers will be asked for to the questions posed, along with the offer of space for the making of a 'position statement' covering aspects which the questions do not touch on. Hopefully the responses to the questions (posed by the editor, associate editors Dr Craig Liebenson DC, Judith Walker Delany, as well as Dr Zachary Comeaux DO and with assistance from Dr Peter Gibbons DO) will provoke additional feedback from readers, a selection of which will be published in subsequent issues of JBMT.

Questions

1. Does the poor inter-observer reliability of palpation methods make you question the validity and usefulness of an examination based upon this skill? If not why not?
2. How do you think palpation and clinical assessment should be taught/studied so that its validity and clinical potential can be best demonstrated?
3. Should we depend less on palpation and assessment methods in clinical settings, since their reliability seems to be so poor?
4. What other tests do you use so that your clinical examination's reliability can be bolstered?
5. If treatment based on possibly unreliable assessment and palpation methods is apparently effective, what does this say about the value of the methods being used? (for example: is apparently successful treatment, based on apparently unreliable assessment methods, valuable mainly for its placebo effects?)
6. Should palpatory diagnostic findings be accepted as having a subjective/interpretative value similar to interpretation of radiographic findings, or other laboratory data, and therefore be capable of being integrated into a treatment plan by a skilled therapist?

REFERENCES

| | | |
|--|--|--|
| <p>Bogduk N 1997 Clinical anatomy of the lumbar spine and sacrum, 3rd edn. Churchill Livingstone, Edinburgh</p> <p>Christensen FB, Laursen M, Gelineck J, Eiskjaer SP, Thomsen K, Bunge CE 2001 Interobserver and intraobserver agreement of radiograph interpretation with and without pedicle screw implants: the need for a detailed classification system in posterolateral spinal fusion.</p> | <p>Spine March 1; 26: 538-543; discussion 543-544</p> <p>Fleiss J 1981 Statistical Methods for rates and proportions. Wiley, New York</p> <p>Jull G, Bogduk N 1988 Accuracy of manual diagnosis for cervical zygapophysial joints. Med J Australia 148: 233-236</p> <p>Lebouef-Yde C, Kyvik K 2000 Is it possible to differentiate people with or without low-back pain on the basis of tests of lumbo-pelvic dysfunction? Journal of Manipulative and</p> | <p>Physiological Therapeutics 23: 160-167</p> <p>Lee D 1999 The pelvic girdle. Churchill Livingstone, Edinburgh</p> <p>Lord S, Bogduk 1996 N Cervical synovial joints as sources of post-traumatic headache. In: Allen M (ed) Musculoskeletal Pain Emanating from Head and Neck. Haworth Medical Press, New York</p> <p>McKenzie C, Taylor N 1997 Can Physiotherapists locate lumbar spinal levels of palpation. Physiotherapy 83: 235-239</p> |
|--|--|--|

Meijne W, van Neerbos K, Aufdemkampe G, van der Wurff P 1999 Intraexaminer and Inter-examiner reliability of the Gillet test. *Journal of Manipulative and Physiological Therapeutics* 22: 4–9
O'Haire F, Gibbons P 2000 Inter-examiner and intra-examiner agreement for

assessing sacroiliac anatomical landmarks using palpation and observation: pilot study. *Manual therapy* 5: 13–20
Slipman C, Sterenfeld E, Chou L, Herzog R, Vresilovic E 1998 The predictive value of provocative SIJ stress manoeuvres in the

diagnosis of sacroiliac syndrome. *Archives of Physical Medicine and Rehabilitation* 79: 288–292
Vincent-Smith B, Gibbons P 1999 Inter-examiner and intra-examiner reliability of standing flexion test. *Manual therapy* 4: 87–93