International Journal of Osteopathic Medicine 25 (2017) 21–29

Contents lists available at ScienceDirect

International Journal of Osteopathic Medicine

journal homepage: www.elsevier.com/ijos

Critical Review

Diagnostic reliability of osteopathic tests: A systematic review

Fabio Basile [a,*], Roberto Scionti [a], Marco Petracca [a,b]

[a] Centre pour l'Etude, la Recherche et la Diffusion Osteopathiques (CERDO), Via Magliano Sabina, 23, 00199 Rome, Italy
[b] San Pietro Fatebenefratelli Hospital, Rome, Italy

ARTICLE INFO

Article history:
Received 12 May 2015
Received in revised form
1 March 2017
Accepted 6 March 2017

Keywords:
Osteopathic medicine
Reliability
Systematic review
Somatic dysfunction

ABSTRACT

Objective: The objective of this systematic review is to identify, evaluate, and synthesize the available evidence concerning studies on diagnostic reliability of osteopathic tests.
Data sources: A structured literature search was conducted of medical databases Cochrane library, MedLine, EMBASE, CINAHL from 1982 to December 2014.
Study selection: Inclusion criteria were journal articles on reliability of osteopathic diagnostic palpatory tests. Studies were excluded if the examiners were not osteopaths or if the studies did not examine the reliability of diagnostic methods.
Data extraction: The included studies were appraised by the three authors using an 11-items scale instrument (QAREL) developed to assess the quality of studies of diagnostic reliability. The search strategy identified 159 studies. However, only 17 met the inclusion criteria.
Data synthesis: The included studies involved a total of 406 patients (64 symptomatic) and 2 studies used anatomical models. 278 examiners were involved, which included 73 osteopaths and 205 osteopathic students. The body sites studied were divided between those of the pelvis, the lower limbs, the spine, and the cranium. Results of the QAREL checklists indicated that most domains met the criteria for being of high quality.
Conclusions: The quality of the reviewed studies was good, while the levels of diagnostic reliability were heterogeneous. Intra-examiner reliability was higher than inter-examiner reliability. There were no significant differences in respect to evaluated osteopathic clinical diagnostic tests and body sites, either in degrees of palpatory pressure, nor in examiners' experience. However, proper study designs, consensus training and standardizing procedures, such as the developing of models, could be effective in improving the reliability of palpatory tests.

1. Introduction

Osteopathy is a primary contact healthcare profession in most countries where it is practised. It diagnoses and treats dysfunctions in the mobility of bodily tissues which affect the state of health [1,2]. It is a concept of health care supported by expanding scientific knowledge that embraces the concept of the unity of the living organism's structure (anatomy) and function (physiology) [3].

Somatic dysfunction is an impaired or altered function of related components of the somatic (body framework) system; skeletal, arthrodial and myofascial structures and their related vascular, lymphatic and neural elements. It is diagnosed by osteopathic clinicians using palpation, a highly trained and practiced sense of touch, which is a hallmark of the osteopathic medicine [1]. Osteopaths follow specific criteria identified by the mnemonic T-A-R-T (Tissue texture changes – Asymmetry – Restriction of motion – Tenderness), and treat somatic dysfunctions using osteopathic manipulative treatment (OMT) [4].

Thus, in order to determine the areas requiring treatment the osteopath uses diagnostic tests. The fundamental requirements of a diagnostic test are that it be both accurate and reliable. A diagnostic test should be reproducible in the same patient by two or more independent examiners, or at the very least, by the same examiner on two separate occasions. These are referred to as inter-examiner and intra-examiner reliability. In other words, the diagnostic reliability is the reproduction of the results when a test is repeated to evaluate an unchanged attribute. An unreliable test will not give reliable results and is not suitable to base decisions about diagnosis and treatment [5].

* Corresponding author. Via Tenente Ugo Granafei, 50, 72023 Mesagne, BR, Italy.
E-mail addresses: fabiobasile.osteo@gmail.com (F. Basile), sciontiroberto@libero.it (R. Scionti), marco.petracca1@hotmail.it (M. Petracca).

http://dx.doi.org/10.1016/j.ijosm.2017.03.004
1746-0689/

The type of statistical test used for analysis in diagnostic studies depends on the type of data generated by the diagnostic test. For continuous data, such as joint range of motion or blood pressure, the intraclass correlation coefficient (ICC) is used [6]. For diagnostic tests that generate categorical data, such as ‘positive’ or ‘negative’, or “mild, “moderate”, or “severe”, the kappa statistic is used [7]. In osteopathy, many diagnostic tests generate categorical data, and in many instances continuous data are reduced into categories for ease of reporting [5].

Previous reviews on reliability of clinical tests have included studies in which examiners were not exclusively osteopaths. Thus, the diagnostic tests were performed by other health care professionals, such as chiropractors, physiotherapists or manual therapists.

Haneline et al. [8] reviewed 29 studies made on inter-examiner reliability of localizing painful or tender points, landmarks, and position or alignment of bone structures of the spine and sacroiliac joints. They concluded that the reliability for these three methods of static palpation is generally low. Thus, even though many pain palpation studies reported acceptable k levels, no one method of static palpation resulted clearly reliable [8]. The review by Seffinger et al. [9] concluded that pain provocation tests are most reliable and soft tissue paraspinal palpatory diagnostic tests are not reliable [9]. Finally, Stovall et al. [10] reviewed studies investigating the reliability of bony anatomical landmark positional asymmetry assessment in the lumbar spine and pelvis. They concluded that the average inter-examiner reliability for bony anatomical landmark positional asymmetry assessment was slightly above chance for all landmarks except medial malleolus, which had fair reliability [10]. Both Stovall et al. [10] and Seffinger et al. [9] concurred that the intra-examiner reliability is greater than the inter-examiner reliability [9,10].

Many of those within and outside the profession have called for a larger evidence base for osteopathic medicine. Thus, it has become essential for research to focus on the investigation of the diagnostic tests, the assessment and pathophysiology of somatic dysfunction, and the indications and physiological effects of osteopathic treatment. Up to now, no review has been published on diagnostic reliability of studies where the examiners were exclusively osteopathic physicians or osteopathic students. Osteopaths place a large emphasis on palpation during training and practice and may therefore have better reliability for palpation than other manual therapists. Therefore, the objective of this systematic review is to identify, evaluate, and synthesize the available evidence concerning studies on diagnostic reliability of osteopathic tests performed by osteopaths.

2. Materials and methods

The research team conducted the systematic review between March through April 2014, with an update in December 2014. That included 3 practitioners with experience in in osteopathy and database searches, clinical research, evidence-based medicine, research design, and statistics methodology.

A search was conducted of the Cochrane library, MEDLINE-PubMed, EMBASE, and Cumulative Index to Nursing and Allied Health Literature (CINAHL) databases for the years 1982 to December 2014. This search involved the following terms: “osteopath*”AND “reliability” in the title/abstract search field.

2.1. Inclusion and exclusion criteria

To be included in the review, articles had to be original studies that investigated osteopathy and reliability of the osteopathic tests, including measurement for the intra- and/or inter-examiner reliability in the English language. Diagnostic tests had to be performed exclusively by osteopaths or osteopathic students. Studies were excluded if they did not examine diagnostic reliability, or were performed by examiners who were not osteopaths.

2.2. Quality assessment

The studies included in this review were appraised for quality by each of the 3 authors using the Quality Appraisal of Reliability Studies (QAREL), an 11-items scale instrument made by Lucas et al., [11] which was developed as a specific quality appraisal tool for studies of diagnostic reliability. It is composed of 21 questions, provided in a data extraction form, summarized in a 11-items checklist. Each item was considered separately for its impact on the quality of the study rather than relying on an overall quality score, therefore QAREL did not use an overall numeric scoring system. The 11-items on the checklist can be answered with 4 categorical responses:“yes”, “no”, “unclear” and, in addition, some items include the option “not applicable”. Yes response indicates a good quality aspect of the study, whereas a No response indicates a poor quality aspect. Unclear is given to unsure or insufficient information, and N/A to uncorrelated information. The 11 points explore 7 domains: items 1 and 2 consider sampling bias and the representativeness of subjects and raters, items 3–7 relate to rater blinding, item 8 relates to the order in which raters or subjects were examined, item 9 relates to the suitability of the time interval among repeated measurements, item 10 considers whether the test was applied and interpreted appropriately and item 11 considers the statistical analysis of reliability [11]. Disagreements about the evaluation of the items' quality were resolved by consensus.

2.3. Data analysis

Characteristics of the studies, findings such as ranges and means of k values, and degrees of reliability were summarized in a table to permit comparison of the outcomes. The examiner reliability for the included studies was characterized using the k statistic for categorical data or the intraclass correlation coefficient (ICC) for continuous data. One study used the 95% limits of agreement between examiners [30] and another study used the Fisher's exact test [17]. Landis and Koch [7] have proposed the following as standards for strength of agreement for the kappa coefficient: 0.00, none; 0.00 to 0.20, slight; 0.21 to 0.40, fair; 0.41 to 0.60, moderate; 0.61 to 0.80, substantial; 0.81 to 1.0, almost perfect [7]. Studies reporting ranges of k values 0.4 or greater were considered clinically acceptable by this review. Fleiss [6] provided a general guideline for ICC interpretation: less than 0.40, poor reliability; 0.40 to 0.75, fair to good reliability; greater than 0.75, good reliability [6].

3. Results

The search generated 159 citations, although, after reading the full-text articles, 85 articles did not meet the inclusion criteria and were excluded. Thus, 22 articles were considered to be potentially relevant, however, another 5 were excluded for the following reasons: reliability of an instrument-based measure [12–14], non-osteopathic examiners [15,16]. Thus, a total number of 17 studies met our inclusion criteria and were included in this systematic review (Fig. 1).

The included studies involved a total number of 406 patients, of whom 64 were symptomatic. Two studies used anatomical pelvis models. Furthermore 278 examiners, of whom 73 osteopaths and 205 osteopathic students, were comprised. The studies originated the US [17–19,23,33], Australia [20–22,24–27,31], UK [28,29,32] and Austria [30]. The region of the body investigated were mostly

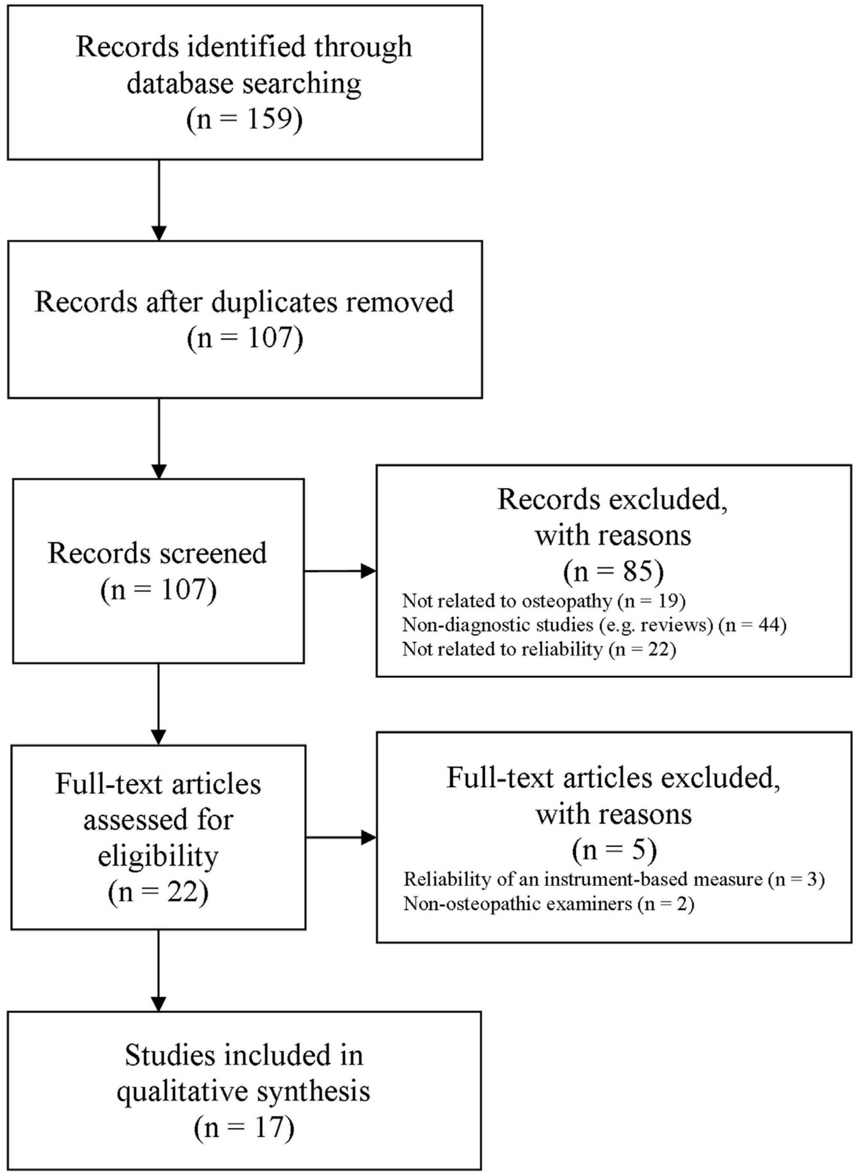

Fig. 1. Flow chart.

divided between those of the pelvis, the lower limbs and the spine: 1 study of the pelvis and spine [29], 7 studies of the pelvis [17,20,22,24,26,32,33], of which 2 included the lower limbs [20,24], 1 of the lower limbs [21] and 5 of the spine [18,19,27,28,31], of which 3 studies included the lumbar spine [18,19,31], 1 the thoracic spine [27] and 1 in which both were included [28]. Finally, 3 studies on the cranium were included [23,25,30], of which 2 examined palpation on the cranium and sacrum together [25,30]. The 3 studies in the cranial field focused on palpation of the cranial rhythmic impulse (CRI) rate [23,25,30]. One of these 3 studies assessed the cranial strain patterns [23] and 2 investigated the "core-link" theory of craniosacral interaction [25,30]. The static and motion palpation tests, examined in the studies selected by the review, are based on the criteria identified by the mnemonic TART [4], as commonly adopted in osteopathic medicine. Tissue texture change was examined in 3 studies [18,19,27], asymmetry in 12 studies [17–22,24,26,29,31–33], restriction of motion in 4 studies [19,23,28,29] and tenderness was assessed only in 2 studies of this systematic review [18,19]. The k statistic was used in 14 studies [17–24,26,27,29,31–33] and the intraclass correlation coefficient (ICC) was used in 2 studies [25,28]. Descriptions of the characteristics of the studies are summarized in Table 1.

Osteopathic tests showed inconsistent diagnostic reliability. Of the 17 studies, 4 reported acceptable k values [18,19,21,23] and 1 reported acceptable ICC values [28]. One study of the anatomical pelvis model reported acceptable k value [33].

Intra-examiner reliability was generally found to be higher than inter-examiner reliability. Intra-examiner reliability of body sites, tested by means of TART parameters in 11 of the 17 studies included in this review, resulted variable. Intra-examiner reliability for palpating tissue texture changes were not measured. Intra-examiner reliability for palpating asymmetry was slight [31] ($k = 0.04$) for lumbar spine, and ranged from: none [24] to moderate [20] ($k = -0.29$ to 0.54) for PSIS (pelvis), fair [24] to substantial [20] ($k = 0.25$ to 0.65) for ASIS (pelvis), none to moderate [24] ($k = -0.29$ to 0.44) for SILA (pelvis), none to almost perfect [24] ($k = -0.28$ to 0.83) for sacral sulci (pelvis), none to almost perfect [24] ($k = 0.20$ to 1.00) for medial malleoli (lower limb). Cranial and

Table 1
Study characteristics and reliability (main findings, degree).

Study	TART[a]	Region	Subjects	Examiners' experience	Study type	Main findings[b]	Degree of reliability[c]
Bengaard K et al. (2012) [17] *(see below for model's results)*	Asymmetry	pelvis	12	44 DO 95 St	Inter	$k = 0.028$ (ASIS: St and DO) $k = 0.038$ (ASIS: St)	slight
Degenhardt BF et al. (2005) [18]	Tissue texture changes Asymmetry Tenderness	lumbar spine (L1-L4)	119 Asx	3 DO, $\leq$ 10 y	Inter	$k = 0.45$ (Texture changes)[d] $k = 0.20$ (Asymmetry: Ant/Post)[d] $k = 0.34$ (Asymmetry: transverse plan)[d] $k = 0.68$ (Tenderness)[d]	slight to substantial
Degenhardt BF et al. (2010) [19]	Tissue texture changes Asymmetry Restriction of motion Tenderness	lumbar spine (L1-L4)	64 (48 Asx, 16 Sx)	2 DO, 3–10 y	Inter	$k = 0.23$ to 0.55 (Texture changes)[d] $k = 0.56$ to 0.72 (Asymmetry)[d] $k = 0.30$ to 0.50 (Restriction of motion)[d] $k = 0.61$ to 0.88 (Tenderness)[d]	fair to almost perfect
Fryer G et al. (2005) [20]	Asymmetry	pelvis - legs	10 Asx	10 St, final-year	Intra	$k = 0.07$ (SILA-Posterior/Ant) $k = 0.20$ (SILA-Inferior/Sup) $k = 0.41$ (seated flexion test)[d] $k = 0.54$ (PSIS)[d] $k = 0.65$ (ASIS)[d] $k = 0.68$ (medial malleoli)[d]	slight to substantial
					Inter	$k = 0.04$ (SILA-Posterior/Ant) $k = 0.04$ (SILA-Inferior/Sup) $k = 0.08$ (PSIS) $k = 0.14$ (seated flexion test)[d] $k = 0.24$ (ASIS)[d] $k = 0.31$ (medial malleoli)[d]	slight to fair
Fryer G et al. (2006) [21]	Asymmetry	legs	10 Asx	5 St final-year	Intra	$k = 0.31$ (medial malleoli: subjects NOT selected) $k = 0.94$ (medial malleoli: subjects selected)	fair to almost perfect
					Inter	$k = 0.22$ (medial malleoli: subjects NOT selected) $k = 0.94$ (medial malleoli: subjects selected)	fair to almost perfect
Gibbons P et al. (2002) [22]	Asymmetry	pelvis	27 Asx	8 St, final-year	Intra	$k = 0.35$ (pelvis lankmarks: no intervention) $k = 0.34$ (pelvis lankmarks: 0.5 cm heel insert) $k = 0.14$ (pelvis lankmarks: 1 cm heel insert)	slight to fair
					Inter	$k = 0.27$ (pelvis lankmarks: no intervention) $k = 0.28$ (pelvis lankmarks: 0.5 cm heel insert) $k = 0.19$ (pelvis lankmarks: 1 cm heel insert)	slight to fair
Halma KD et al. (2008) [23]	Restriction of motion	cranium	48 (15 Asx, 33 Sx)	2 DO, 6–14 y	Intra	$k = 0.23$ (cranial rhythmic impulse) $k = 0.67$ (cranial strain patterns) $k = 0.52$ (left anterior quadrant) $k = 0.44$ (right anterior quadrant) $k = 0.33$ (left posterior quadrant) $k = 0.50$ (right posterior quadrant)	fair to substantial
Kmita A et al. (2008) [24]	Asymmetry	pelvis - legs	9 (4 Asx, 5 Sx)	2 DO, Exp 2 St, final-year	Intra	$k = -0.29$ to 0.39 (PSIS) $k = -0.28$ to 0.83 (sacral sulcus) $k = -0.29$ to 0.44 (SILA-Posterior/Ant)	none to almost perfect

Table 1 (*continued*)

Study	TART[a]	Region	Subjects	Examiners' experience	Study type	Main findings[b]	Degree of reliability[c]
						k = −0.29 to 0.34 (SILA-Inferior/Sup) *k* = 0.25 to 0.63 (ASIS) *k* = 0.20 to 1.0 (medial malleoli)	
					Inter	*k* = 0.04 (PSIS) *k* = −0.4 (sacral sulcus) *k* = −0.03 (SILA-Posterior/Ant) *k* = −0.01 (SILA-Inferior/Sup) *k* = 0.13 (ASIS) *k* = 0.17 (medial malleoli)	none to slight
Moran RW et al. (2001) [25]	*none*	cranium - sacrum	11 Asx	2 DO, 4.5–6.5 y	Intra	*ICC* = −1.47 to −1.46 (cranium-sacrum) *ICC* = 0.52 to 0.65 (sacrum) *ICC* = 0.47 to 0.73 (cranium)	none to substantial
					Inter	*ICC* = −0.09 to 0.31 (cranium-sacrum) *ICC* = −0.02 (sacrum) *ICC* = 0.05 (cranium)	none to fair
O'Haire C et al. (2000) [26]	Asymmetry	pelvis	10 Asx	10 St, fifth year	Intra	*k* = 0.21 (SILA) *k* = 0.24 (sacral sulcus) *k* = 0.33 (PSIS)	fair
					Inter	*k* = 0.04 (PSIS) *k* = 0.07 (sacral sulcus) *k* = 0.08 (SILA)	slight
Paulet T et al. (2009) [27]	Tissue texture changes	thoracic spine	10 Sx	10 St, 4 final-year	Inter	*k* = 0.26 (total assessment: n.10) *k* = 0.32 (first 5 assessment)	fair
Potter L et al. (2006) [28]	Restriction of motion	lumbar - thoracic spine	12 Asx	1 DO, 9 y	Intra	*ICC* = 0.70 (thoracic spine) *ICC* = 0.96 (lumbar spine)	fair to excellent
Rajendran D et al. (2011) [29]	Asymmetry Restriction of motion	pelvis - lumbar spine	4 Asx	20 St, third year	Inter	*k* = −0.0476 (SILA) *k* = −0.0425 (backwards bending test) *k* = 0.0170 (lumbar spring test) *k* = 0.0330 (sacral sulcus)	none to slight
Sommerfeld P et al. (2004) [30]	*none*	cranium - sacrum	49 Asx	2 DO, 7 y	Intra Inter	*95% limits of agreement*[e]	*not reliable*
Spring F et al. (2001) [31]	Asymmetry	lumbar spine (L4)	10 Asx	10 St, final-year	Intra Inter	*k* = 0.04 (Asymmetry) *k* = 0.04 (Asymmetry)	slight slight
Sutton C et al. (2013) [32]	Asymmetry	pelvis	1 Asx	10 DO, 5–25 y 15 St, third year 15 St, final-year	Inter	*k* = 0.025 (PSIS, St - third year) *k* = 0.058 (PSIS, DO) *k* = 0.065 (PSIS, St - final year)	slight
Stovall BA et al. (2010) [33] [g]	Asymmetry	pelvis	1 pelvic model	5 DO, 5–40 y 5 St, final-year	Intra	*k* = 0.49 (DO) *k* = 0.52 to 0.59 (St)[d]	moderate
					Inter	*k* = 0.46 (DO) *k* = 0.37 to 0.48 (St)[d]	fair to moderate
Bengaard K et al. (2012) [17] [g]	Asymmetry	pelvis	1 pelvic model	44 DO 95 St	Inter	*Fisher's exact test*[f]	*not reliable*

Asx = asymptomatic; Sx = symptomatic; DO = doctor of osteopathy; St = osteopathic student; Exp = experienced; Intra = intraexaminer reliability; Inter = interexaminer reliability.
DESCRIPTION: This table summarizes the studies' characteristics (such as TART parameters assessed, region of the body evaluated, sample of subjects, sample of examiners plus qualification and experience in years, type of reliability investigated), main findings (such as ranges or means of κ/ICC values), and degrees of reliability according to Landis and Kock or Fleiss classification.
[a] T-A-R-T (Tissue texture changes – Asymmetry – Restriction of motion – Tenderness) parameters, measured by examiners.
[b] Range or mean values shown in ascending order, details shown in brackets (...)
[c] k Landis and Koch, ICC Fleiss.
[d] Finding increased after consensus training.
[e] 95% limits of agreement Bland and Altmann.
[f] Fisher's exact test, one-tailed $p < .0001$.
[g] Studies where one anatomical model was used.

thoracic asymmetries were not measured. Intra-examiner reliability for palpating restriction of motion was: fair [28] (ICC = 0.70) for thoracic spine and excellent [28] (ICC = 0.96) for lumbar spine. Halma et al. [23] assessed restriction of motion of the 4 cranial quadrants: the overall intra-examiner reliability ranged from fair to moderate (k = 0.33 to 0.52). Restrictions of motion of pelvis and lower limb were not measured. Intra-examiner reliability for palpating tenderness was not measured. The intra-examiner reliability of the two pelvic models was measured only by Stovall et al. [33] for palpating asymmetry and findings ranged from $k = 0.33$ to 0.52, a moderate degree of reliability.

Inter-examiner reliability was assessed in 15 of the 17 studies included. Inter-examiner reliability for palpating tissue texture changes ranged from: fair to moderate [19] ($k = 0.23$ to 0.55) for lumbar spine, fair [27] ($k = 0.26$ to 0.32) for thoracic spine and no other body sites were measured for tissue texture changes. Inter-examiner reliability for palpating asymmetry ranged from: slight [31] to substantial [19] ($k = 0.04$ to 0.72) for lumbar spine, slight [32] to fair [22] ($k = 0.025$ to 0.28) for PSIS (pelvis), slight [17] to fair [20] ($k = 0.028$ to 0.24) for ASIS (pelvis), none [29] to slight [26] ($k = -0.0476$ to 0.08) for SILA (pelvis), none [24] to slight [26] ($k = -0.4$ to 0.07) for sacral sulci (pelvis), slight [24] to almost perfect [21] ($k = 0.17$ to 0.94) for medial malleoli (lower limb). Asymmetries of thoracic spine and cranium were not measured. Inter-examiner reliability for palpating restriction of motion ranged from slight [29] to moderate [19] ($k = 0.017$ to 0.50) for lumbar spine, no other body sites were measured for restrictions of motion. Inter-examiner reliability for palpating tenderness ranged from substantial to almost perfect [19] ($k = 0.61$ to 0.88) for lumbar spine and no other body sites were measured for tenderness. The inter-examiner reliability of the two pelvic models was reported only by Stovall et al. [33] for palpating asymmetry and findings ranged from fair to moderate ($k = 0.37$ to 0.48). Bengaard et al. [17] did not report any k values in regards to inter-examiner reliability of their pelvic model, because Fisher's exact test was preferred.

Results of the other studies were included into the above-mentioned ranges but resulted variable or below the acceptable k or ICC values (Table 1).

The QAREL checklists of the 17 studies indicated that most domains met the criteria for being of high quality, with a majority of Yes replies (61%) (Table 2). In addition to the quality assessed by QAREL, the present systematic review analyzed the results for each of the 11 items. Thus, the sample of subjects (*item 1*) was generally a good aspect, however some studies reported inconsistent results. The sample of raters (*item 2*), such as osteopaths, were rated as being good quality for all 17 studies. Conversely, rater blinding (item 3 to 7) was the most conflicting area. *Item 3*, inter-examiner blinding, was rated as being good quality; on the other hand, the intra-examiner blinding (*item 4*) had a mix of high and low quality ratings. *Item 5*, concerning the blinding to the accepted reference standard, was always rated as "not applicable". Furthermore, blinding protocols in regards to the clinical information of the subjects (*item 6*) was conflicting: there was insufficient information in more than half of the studies. *Item 7* was rated with the largest number of "no" responses. The order in which raters or subjects were examined (*item 8*) was generally found to be a good aspect. *Item 9* received the largest number of "unclear" and *Item 10* was considered as a good aspect. Finally, the statistical analysis of reliability (*item 11*) was rated as "yes" for all the studies reviewed.

Table 2
4-answers summary (QAREL).

Study	Yes	No	Unclear	N/A
Bengaard K et al. (2012)[17a]	3	2	4	2
Degenhardt BF et al. (2005) [18]	6	1	2	2
Degenhardt BF et al. (2010) [19]	7		2	2
Fryer G et al. (2005) [20]	8	1	1	1
Fryer G et al. (2006) [21]	8		2	1
Gibbons P et al. (2002) [22]	6	1	3	1
Halma KD et al. (2008) [23]	8		2	1
Kmita A et al. (2008) [24]	10			1
Moran RW et al. (2001) [25]	6	2	2	1
O'Haire C et al. (2000) [26]	8	1	1	1
Paulet T et al. (2009) [27]	5	2	2	2
Potter L et al. (2006) [28]	5		4	2
Rajendran D et al. (2011) [29]	7	1	1	2
Sommerfeld P et al. (2004) [30]	5	1	4	1
Spring F et al. (2001) [31]	7		3	1
Sutton C et al. (2013) [32]	6	1	2	2
Stovall BA et al. (2010)[33a]	9			2
Total	**114**	**13**	**35**	**25**

DESCRIPTION: This table summarizes the 4 answers of the 11-items scale of the QAREL checklists, showing the total amount of each answer.
[a] Studies where one anatomical model was used.

4. Discussion

The studies included in this systematic review were judged to be of good quality, as assessed using the QAREL tool. Conversely, further analysis of the 17 studies showed an inconsistent diagnostic reliability, providing insight into strengths and weaknesses of their methodologies and study designs, such as examiners' experience and training, use of novel models and stability of the variables. Furthermore, reliability for each of the different palpatory test according to the TART mnemonic, and the different body sites, was examined.

Tissue texture changes were examined in 3 inter-examiner reliability studies [18,19,27] and improved from fair to moderate after a consensus training [18,19], but decreased after repeated assessments [27]. Because so few studies have examined the reliability of tissue texture change assessment between examiners, further studies are required to determine the reliability of this clinical finding. Furthermore, these studies were based only on the spine examination and there are no studies on other body sites or intra-examiner reliability.

Asymmetry was measured in most of the intra- and inter-examiners reliability studies [17–22,24,26,29,31–33], with a prevalence of pelvis and lower limbs studies. Some studies assessed asymmetry after the subjects changed position [18,20,31], but most of the studies examined static segmental positional asymmetry without any movement of the test subjects [17,19–22,24,26,29,32,33] to maintain the stability of the variables. Intra-examiner reliability was higher than inter-examiner, but the findings were inconsistent and prevent them from being considered clinically useful. Effectively, few studies reported acceptable intra-examiner k values for medial malleoli [21,24] and ASIS [20] asymmetry, but the overall kappa scores of the reviewed studies varied considerably, indicating that some examiners achieved reasonably high agreement, whereas others did not. Moreover, those results were not consistent in respect to the other body sites. A possible explanation for higher reliability of the studies based on asymmetry of pelvis landmarks and medial malleoli could be that the landmarks in these regions were more easily detected because there was less obstructing fat and muscle. In fact, some authors found inconsistency when assessing asymmetry of other body sites, such as the spine, because of the tissue hypertonicity and thickness [19]. Furthermore, reliability for assessing symmetry in participants improved after a consensus training [18,19]. The osteopathic profession should explore ways to make these assessments more reliable and examine the validity of landmark asymmetry.

Restriction of motion was tested in 4 studies [19,23,28,29] based on spinal or cranial restrictions, equally divided between intra- [23,28] and inter-examiner reliability [19,29]. However, findings could not be compared between the spinal segments and the cranial bones because articular mobility, motion-testing and methodologies are different. Intra-examiner reliability within the cranial fields resulted good when diagnosing cranial strain patterns [23]. This reliability was also good when diagnosing lumbar spine restrictions, however, intra-examiner reliability of thoracic spine motion palpation was poor [28]. Inter-examiner reliability for restriction of motion assessment was lower than intra-examiner, but improved when assessing lumbar dysfunctions after consensus training [19]. Future studies on reliability of restriction of motion tests are necessary.

Tenderness was the last TART parameter investigated. It is a response by the patient to palpation by the physician, so this parameter is the most likely subjective finding. Even if some authors defined it as the presence of pain elicited by palpation [2,4], tenderness is often distinguished from pain because of pressure during examination and various subjective sensations reported by patients. Degenhardt et al. [18,19] used pain provocation test to assess tenderness of the lumbar spine, which improved after consensus training to provide acceptable reliability between examiners. Further investigations need to evaluate the intra-examiner reliability of tenderness tests, evaluate other body sites and better understand the stability of variables.

The above-mentioned TART parameters were analyzed to understand the reliability of assessing the different body sites. Reliability of the osteopathic tests for the 4 TART findings on the lumbar spine was good in some studies [18,19,28]. Pelvis and lower limbs were mostly analyzed for asymmetry, showing some good intra-examiner reliability [20,21]. Assessment of the cranium was only investigated for intra-examiner reliability to assess restriction of motion, and findings were acceptable [23]. All studies of the cranial region evaluated the CRI rate and the craniosacral mechanism, rather than the TART parameters. However, these studies failed to support the CRI rate palpation and the “core-link” hypothesis [25,30]. In future studies, it is important to consider increasing the number of reliability studies focused on TART parameters of the main body sites to determine the extent to which reliability of somatic dysfunction evaluation could be generalized in the clinical setting.

Reliability and TART parameters could be influenced by some factors, such as consensus training, examiner's experience, as highlighted above, and use of novel models and standardizing procedure. Consensus training concerns discussions about findings, followed by retesting until greater agreement between examiners could be achieved [26]. Degenhardt et al. [18,19] defined a detailed protocol for each inter-observer test during a previous comprehensive consensus training period. When findings were inconsistent between the examiners, focused consensus training was performed as a means of recalibration to understand why assessments were inconsistent. It appears that consensus training improved the reliability of common osteopathic palpatory tests of the lumbar spine [18,19]. As previously mentioned for palpation of asymmetry, this finding is also supported by Stovall et al. [33] for pelvic model. Fryer et al. [20] concluded that examiners who attended a single consensus training session achieved only a marginal increase in reliability. However, findings were not consistent with Degenhardt et al. [18,19] because of the difference in the amount of consensus training [18–20]. There were no studies about the role of consensus training for the tests in the cranial field. Interestingly, training without consensus showed no improvements in reliability [22,26]. Differences between experience of osteopaths and students in osteopathy were examined by some of the studies included in this systematic review. Experienced examiners are presumed to have advanced palpatory skills gained over years of practice. However, previous studies have not found that less experienced palpators demonstrate poorer agreement than experienced practitioners. This finding is supported by some of the studies included in this systematic review [18,20,24,27,32]. The reason could be a lack of consensus on what constitutes expertise in osteopathy. For example, years in clinical practice and advanced professional certification are not able to differentiate between ‘expert’ and ‘average’ practitioners when using clinical outcomes to identify expert physical therapists [32]. A novel pelvic model, developed by Stovall et al., [33] was used to evaluate improvements in assessing inter-examiner and intra-examiner reliability. The development of realistic models for training of palpatory diagnosis, along with increased understanding of anatomical and physiologic models of function and dysfunction, is needed to improve training and education in all professions in manipulative medicine [33]. Although the reliability of assessment of pelvic model asymmetry was acceptable, a fixed model cannot likely represent the complexity of palpating human (real) somatic dysfunctions. Beengaard et al. [17] assessed palpation of symmetry of two paired structures, such as the anterior superior iliac spine (ASIS), using a fixed plastic model but found low concordance with assessments and the actual settings on the models [17]. Sutton et al. [32] and Gibbons et al. [22] suggested an idea to standardize the palpation for symmetry on human bodies: a heel wedge was used to alter the height of the pelvic landmarks which was hidden from the examiners. However, the clinical tests used to detect sacroiliac joint dysfunction and leg length inequality have not shown good reliability, because the degree of asymmetry prior to insertion was not defined [32] and the wedge sizes were not large enough (5–10 mm) to produce a reliable detection of discrepancy [22,32]. Potter et al. [28] used a standardized UV marking pen assessment protocol to identify a joint exhibiting signs of segmental dysfunction in the thoracic spine. This method removed the possible measurement error inherent in needing to name or label the correct spinal segment [28]. Future studies need to evaluate the role of standardizing models and procedures to better understand whether calibration and standardization could improve reliability of diagnostic tests. For example, osteopathic students could frequently train palpation and reliability on different realistic models with standardized asymmetries, even if these are different from the osteopathic practice on human bodies. Detection of asymmetry on a model could be measured easier than human and it could be a way to quickly understand the levels of diagnostic reliability within the practitioners. Comprehensive palpation training with both teaching models and humans, standardized procedures and assessment protocols, may improve the diagnostic reliability of osteopathic tests used by osteopath graduates.

This systematic review also highlighted other possible reasons that could lead to inconsistency of reliability, such as subjects' positions and movements, and examiners' palpatory characteristics. Subjects' positions and movements influence the stability of the variables. Spring et al. [31] determined the inter- and intra-examiner reliability of a commonly used three-part positional diagnostic screen for the lumbar spine (L4), according to the TART mnemonic. However, the testing conditions simulated clinical conditions with respect to position changes, a factor that could potentially lead to changes in alignment [31]. Furthermore, even when subjects appeared to remain remarkably still there was little doubt that they would adjust their position. Basically, small movements may lead to instability of palpatory cues and contribute to reduce intra- and inter-examiner agreement [26]. Examiners' degrees of palpatory pressure, associated with repeated evaluations, could contribute to systematic error and a reduction in agreement [17,22,23,26,27]. There is some evidence that palpation pressure biofeedback techniques may improve reliability [22,26]. Effectively, Degenhardt et al. [18,19] are the only two studies where the examiners reproduced the amount of pressure applied on a calibrated scale to quantify the amount of pressure [18,19]. Furthermore, sensory limits or use different surfaces of the fingers could contribute to variability between examiners [26,27] and the part of the bone to palpate, although all the examiners are familiar with the palpation, could not be the same all the time [17,22,27]. For example, Beengaard et al. [17] assumed that the ASIS may be palpated by an examiner hooking his or her thumb under the ASIS protrusion, finding the center of the ASIS protrusion, or perhaps palpating from above. Examiners could be also unaware of which eye is dominant and that could lead to inconsistent outcomes, especially when the measurement is set up from one side of the table [17].

The study design could also affect reliability and this systematic review highlighted the most conflicting domains by means of the use of the QAREL, such as sampling bias and the representativeness of subjects (*item 1*), examiner blinding (*item 3–7*) and the suitability of the time interval among repeated measurements (*item 9*), as follow. The applicability of a study is affected by the selection of subjects. Subjects who represent typical patients on whom the test would be used in practice should be used in the study of clinical tests. For instance, Fryer et al. [21] showed an almost perfect intra-examiner and inter-examiner reliability following subject selection for malleoli asymmetry, which suggests that clinicians can reliably detect medial malleoli asymmetries of greater than approximately 4 mm difference. This study suggests that the amount of asymmetry that exists in the bony landmarks of a subject has a profound influence on the intra-examiner and inter-examiner agreement for assessment of those landmarks [21]. Furthermore, a high BMI, such as obesity, was deemed to be a potential source of error, increasing difficulty in palpation because of the excessive adipose tissue [22,28,32]. Thus, the sample of subjects assessed by the QAREL (*item 1*) reported some inconsistent results about the characteristics (symptomatic or healthy, sex, BMI, etc.) and the sample size, not always adequate for a statistical significance.

A failure to blind raters would compromise the quality of the study and render it inapplicable [11]. Effectively, intra-examiner blinding (*item 4*), blinding to the accepted reference standard (*item 5*), blinding protocols in regards to the clinical information of the subjects (*item 6*) and blinding to additional cues that were not part of the test (*item 7*) received heterogeneous responses due to some unspecified or uncorrelated information about the blinding procedure. Item 5, for example, was specifically rated as “not applicable” because there are no gold standards in osteopathy: a gold standard, or what is now referred to as the reference standard, is the test believed to be the most accurate, or most direct measure, of the index condition by which we could predetermine the prevalence of the signs in the sample, irrespective of the sample size [5,24].

Moreover, time interval between measurements (*item 9*) was another unclear topic: some authors defined 30 s, 1 or 2 min' duration for each trials and a minimum rest period of 1.5–2 min between trials [21,22,24,25], in other studies subjects were asked to lie still on the examination table for up to 30 min, whereas other studies had subjects lying for up to 2.5 h [17,21,26]. Thus, *item 9* received the largest number of “unclear” due to insufficient information about the suitability of the time interval among repeated measurements.

Interestingly, Kmita et al. [24] assumed that only when a dysfunction is consistently present it can be considered as potentially etiological in the production of symptoms. Many biological variables fluctuate over time, and are dynamic rather than static. Only when a reliable method of measuring the variable is developed, can normative data be established from which to determine abnormal deviations. If somatic dysfunction is unstable, inconsistently present, or modifiable with simple assessment techniques, how can it be found, how can it be fixed, and how can its correction be determined? It is reasonable to acknowledge that the body is dynamic, and that no two physical examinations will be identical in the same patient. However, this is true of all physical interactions and is exactly why a test must be reliable in order to be useful [24].

Limitations of this review include the search strategy and databases used, the number and characteristics of the studies, and the inclusion criteria. We searched 4 biomedical databases using assorted search terms and also harvested references from the included articles. Nevertheless, it is possible that some relevant citations may have been missed. The number of studies available for analysis was limited, and these were generally not amenable to direct comparisons because of heterogeneity of their methods, statistical analyses, and regions investigated. We only analyzed reliability of diagnostic studies because they can detect the somatic dysfunction by means of osteopathic tests; this is important for understanding and practicing osteopathy even though it limit the number of reliability studies included. Furthermore, in using a quality assessment instrument, such as QAREL, some quality scoring criteria are more detailed/differentiated than others, which introduces an inherent bias. Scores/assigned weights may be biased toward rigor of research methodology and presentation. Since the quality assessment instrument focused on the internal validity of the studies, the quality scores cannot be extrapolated to measure the studies' significance or impact (in terms of findings, relevance to the discipline).

The results of the current study suggest that teaching institutions need to reflect on their curricula to better calibrate and standardize palpatory diagnostic skills.

5. Conclusion

This systematic review was the first to evaluate studies on diagnostic reliability exclusively of osteopathic tests performed by osteopaths and osteopathic students. The quality of the reviewed studies was good, while the levels of diagnostic reliability were heterogeneous. Intra-examiner reliability was found to be higher than inter-examiner reliability. There were no differences in respect to reliability of the 4 TART parameters and the different body sites, even if static segmental positional asymmetry tests for pelvis and lower limbs were the most investigated. Furthermore, no differences were found for subjects' levels of consciousness, positions and movements, as much as examiners' degrees of palpatory pressure, sensory limits, methods of palpation and experience. Reliability could be affected by the study design, such as sample of subjects, blinding protocols, duration of assessments and time interval between measurements. However, consensus training and developing of standardizing procedure, such as novel models, could be effective in improving the reliability of palpatory tests.

Further research into study design and diagnostic reliability of osteopathic tests for somatic dysfunction and TART parameters are necessary.

Author contribution statement

MP conceived the idea for the study, the design and the critical revision for important intellectual content. FB and RS contributed to the design and planning of the research. FB and RS collected, analyzed the data and wrote the first draft of the manuscript in Italian language. FB interpreted and summarized data for presentations and publications. FB also created all the tables, illustration and figures and translated the full paper in English version. All authors edited and approved the final version of the manuscript.

Funding

No funding was received from any pharmaceutical or equipment companies.

APPENDIX

Data extraction form and checklist can be found, in the online version, at qarel.org.

References

[1] Scope of Osteopathic Practice in Europe, SOPE. European Federation of Osteopaths (EFO), Brussels, June 2010.
[2] Chila A. Foundation of osteopathic medicine. third ed. William & Wilkins; 2011. p. 3.
[3] Franke H, Kranke JD, Fryer G. Osteopathic manipulative treatment for nonspecific low back pain: a systematic review and meta-analysis. BMC Musculoskelet Disord 2014;15:286.
[4] Di Giovanna Eileen L, Stanley Schiowitz, Dowling Dennis J. An osteopathic approach to diagnosis and treatment. third ed. Lippincott Williams & Wilkins; 2005. p. 16.
[5] Lucas N, Bogduk N. Diagnostic reliability in osteopathic medicine. Int J Osteopath Med 2011;14:43–7.
[6] Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley and Sons; 1986. p. 7.
[7] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
[8] Haneline MT, Young M. A review of intraexaminer and interexaminer reliability of static spinal palpation: a literature synthesis. J Manip Physiol Ther 2009 Jun;32(5):379–86.
[9] 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 (Phila Pa 1976) 2004 Oct 1;29(19):E413–25.
[10] Stovall BA, Kumar S. Anatomical landmark asymmetry assessment in the lumbar spine and pelvis: a review of reliability. PM R 2010 Jan;2(1):48–56.
[11] Lucas NP, Macaskill P, Irwig I, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability. J Clin Epidemiol 2010;63: 854–61.
[12] Snider KT, Kribs JW, Snider EJ, Degenhardt BF, Bukowski A, Johnson JC. Reliability of Tuffier's line as an anatomic landmark. Spine (Phila Pa 1976) 2008 Mar 15;33(6):E161–5.
[13] Bockenhauer SE, Chen H, Julliard KN, Weedon J. Measuring thoracic excursion: reliability of the cloth tape measure technique. J Am Osteopath Assoc 2007 May;107(5):191–6.
[14] Shaw KA, Dougherty JJ, Treffer KD, Glaros AG. Establishing the content validity of palpatory examination for the assessment of the lumbar spine using ultrasonography: a pilot study. J Am Osteopath Assoc 2012 Dec;112(12): 775–82.
[15] Tong HC, Heyman OG, Lado DA, Isser MM. Interexaminer reliability of three methods of combining test results to determine side of sacral restriction, sacral base position, and innominate bone position. J Am Osteopath Assoc 2006 Aug;106(8):464–8.
[16] Walker BF, Buchbinder R. Most commonly used methods of detecting spinal subluxation and the preferred term for its description: a survey of chiropractors in Victoria, Australia. J Manip Physiol Ther 1997 Nov-Dec;20(9): 583–9.
[17] Bengaard K, Bogue RJ, Crow WT. Reliability of diagnosis of somatic dysfunction among osteopathic physicians and medical students. Osteopath Fam Physician 2012;4:2–7.
[18] Degenhardt BF, Snider KT, Snider EJ, Johnson JC. Interobserver reliability of osteopathic palpatory diagnostic tests of the lumbar spine: improvements from consensus training. J Am Osteopath Assoc 2005 Oct;105(10):465–73.
[19] Degenhardt BF, Johnson JC, Snider KT, Snider EJ. Maintenance and improvement of interobserver reliability of osteopathic palpatory tests over a 4-month period. J Am Osteopath Assoc 2010 Oct;110(10):579–86.
[20] Fryer G, McPherson HC, O'Keefe P. The effect of training on the inter-examiner and intra-examiner reliability of the seated flexion test and assessment of pelvic anatomical landmarks with palpation. Int J Osteopath Med 2005;8: 131–8.
[21] Fryer G. Factors affecting the intra-examiner and inter-examiner reliability of palpation for supine medial malleoli asymmetry. Int J Osteopath Med 2006;9: 58–65.
[22] Gibbons P, Dumper C, Gosling C. Inter-examiner and intra-examiner agreement for assessing simulated leg length inequality using palpation and observation during a standing assessment. J Osteopath Med 2002;5(2):53–8.
[23] Halma KD, Degenhardt BF, Snider KT, Johnson JC, Flaim MS, Bradshaw D. Intraobserver reliability of cranial strain patterns as evaluated by osteopathic physicians: a pilot study. J Am Osteopath Assoc 2008 Sep;108(9):493–502.
[24] Kmita A, Lucas NP. Reliability of physical examination to assess asymmetry of anatomical landmarks indicative of pelvic somatic dysfunction in subjects with and without low back pain. Int J Osteopath Med 2008;11:16–25.
[25] Moran RW, Gibbons P. Intraexaminer and interexaminer reliability for palpation of the cranial rhythmic impulse at the head and sacrum. J Manip Physiol Ther 2001 Mar-Apr;24(3):183–90.
[26] O'Haire C, Gibbons P. Inter-examiner and intra-examiner agreement for assessing sacroiliac anatomical landmarks using palpation and observation: pilot study. Man Ther 2000 Feb;5(1):13–20.
[27] Paulet T, Fryer G. Inter-examiner reliability of palpation for tissue texture abnormality in the thoracic paraspinal region. Int J Osteopath Med 2009;12: 92–6.
[28] Potter L, McCarthy C, Oldham J. Intraexaminer reliability of identifying a dysfunctional segment in the thoracic and lumbar spine. J Manip Physiol Ther 2006;29:203–7. 2006.
[29] Rajendran D, Gallagher D. The assessment of pelvic landmarks using palpation: a reliability study of undergraduate students. Int J Osteopath Med 2011;14:57–60.
[30] Sommerfeld P, Kaider A, Klein P. Inter- and intraexaminer reliability in palpation of the “primary respiratory mechanism” within the “cranial concept”. Man Ther 2004 Feb;9(1):22–9.
[31] Spring F, Gibbons P, Tehan P. Intra-examiner and inter-examiner reliability of a positional diagnostic screen for the lumbar spine. J Osteopath Med 2001;4(2):47–55.
[32] Sutton C, Nono L, Johnston RG, Thomson OP. The effects of experience on the inter-reliability of osteopaths to detect changes in posterior superior iliac spine levels using a hidden heel wedge. J Bodyw Mov Ther 2013 Apr;17(2): 143–50.
[33] Stovall BA, Bae S, Kumar S. Anterior superior iliac spine asymmetry assessment on a novel pelvic model: an investigation of accuracy and reliability. J Manip Physiol Ther 2010 Jun;33(5):378–85.