Reliability of Physical Examination for Diagnosis of Myofascial Trigger Points

A Systematic Review of the Literature

Nicholas Lucas, BSc, MHSc, MPainMed,* † Petra Macaskill, BA, MAppStat, PhD, † Les Irwig, MBCH, PhD, ‡ Robert Moran, BSc, MHSc, ‡ and Nikolai Bogduk, MBBS, PhD, MD, DSc§

Background: Trigger points are promoted as an important cause of musculoskeletal pain. There is no accepted reference standard for the diagnosis of trigger points, and data on the reliability of physical examination for trigger points are conflicting.

Objectives: To systematically review the literature on the reliability of physical examination for the diagnosis of trigger points.

Methods: MEDLINE, EMBASE, and other sources were searched for articles reporting the reliability of physical examination for trigger points. Included studies were evaluated for their quality and applicability, and reliability estimates were extracted and reported.

Results: Nine studies were eligible for inclusion. None satisfied all quality and applicability criteria. No study specifically reported reliability for the identification of the location of active trigger points in the muscles of symptomatic participants. Reliability estimates varied widely for each diagnostic sign, for each muscle, and across each study. Reliability estimates were generally higher for subjective signs such as tenderness (κ range, 0.22–1.0) and pain reproduction (κ range, 0.57–1.00), and lower for objective signs such as the taut band (κ range, −0.08–0.75) and local twitch response (κ range, −0.05–0.57).

Conclusions: No study to date has reported the reliability of trigger point diagnosis according to the currently proposed criteria. On the basis of the limited number of studies available, and significant problems with their design, reporting, statistical integrity, and clinical applicability, physical examination cannot currently be recommended as a reliable test for the diagnosis of trigger points. The reliability of trigger point diagnosis needs to be further investigated with studies of high quality that use current diagnostic criteria in clinically relevant patients.

Key Words: physical examination, trigger point, myofascial pain, reproducibility, reliability, systematic review

(Pain believed to be of musculoskeletal origin is a common complaint in primary care, and is a major public health concern.1–8 In the majority of patients, it is difficult to come to a definitive diagnosis. One explanation for these pain symptoms is that patients have myofascial pain syndrome, a condition characterized by painful trigger points (TPs) in muscles.9–13 Numerous studies report the prevalence of trigger points in various clinical conditions, with one claim that they accounted for up to 85% of patients presenting to a tertiary pain clinic.14 The list of conditions reported to be associated with TPs is extensive, and includes migraine, tension type headache, temporomandibular disorder, neck pain, shoulder pain, epicondylitis, carpal tunnel syndrome, low back pain, pelvic pain, and atypical angina pectoris.10,15–18 Treatments for TPs include injections,11,19 needling,19 spray and stretch,11,13 and massage,11,13 yet the current evidence does not demonstrate that these treatments are more efficacious than placebo.20–23

Pivotal to the appropriate and accurate prescription of any treatment is accurate diagnosis. If the diagnosis is incorrect, an otherwise effective treatment may fail because it is being applied to patients who do not have the condition. Fundamental to accurate diagnosis is the reliability of the test used to make the diagnosis. Reliability is the extent to which examiners, using the same test on the same patients, agree on the results of the test. For the investigation of TPs, various procedures have been employed, such as microdialysis,24 biopsy,25 imaging techniques,26 and electromyography;27 yet none of these are definitive or accepted as a reference standard.10,28,29 Physical examination is the only means by which to establish the diagnosis and consists of firm digital pressure applied to the muscle to identify the diagnostic criteria and elicit feedback from the patient.13

The original diagnostic criteria reported to be essential for the diagnosis of TPs were tenderness within a taut band, a predictable pattern of referred pain with palpation of the taut band, and a painful limited range of movement.13,30 Over time, the criteria for TPs were modified to include reproduction of the patients familiar pain with palpation of the tender point within the taut band, a local twitch response with “snapping” palpation of the taut band, and a “jump sign” when the patient quickly retracts away in response to palpation of the taut band.15,30 Currently,
authorities describe a TP as a hyperirritable nodule located within a taut band of skeletal muscle that when palpated is tender and produces referred pain.\textsuperscript{13,15,25,30} A predictable pattern of pain referral and the local twitch response are each no longer considered to be sufficient or necessary for the diagnosis.\textsuperscript{15,30}

After the identification of a TP, examiners determine if it is “active” (symptomatic) or “latent” (asymptomatic). Essential for the diagnosis of an active TP is the criterion that the patient’s pain can be reproduced by palpation of the tender point within the taut band.\textsuperscript{24} Because only patients with symptoms can have active TPs, it is necessary to demonstrate the reliability of this diagnosis in symptomatic participants. If physical examination were reliable for the identification of TPs, examiners would be able to agree upon the exact location of the taut band, and the presence of a tender nodule within it. If this finding were clinically relevant, then palpation of the nodule within the taut band should also reproduce the patient’s familiar pain.\textsuperscript{15} If a diagnosis requires several diagnostic criteria to be satisfied, such as with TPs, then studies of the reliability of that diagnosis must address agreement on the simultaneous presence of all signs in each patient.

Several studies have been conducted to assess the reliability of palpation for TPs. Given the importance of reliability for the assessment of TPs, the evidence on reliability is worthy of systematic review. The primary objectives of this systematic review were to collect and summarize the available evidence regarding the reliability of physical examination for the diagnosis of TPs, and to analyze studies for their quality and for factors that might affect their applicability in clinical practice. Secondary objectives were to identify deficiencies in the available evidence, and to recommend how these might be redressed. Finally, implications for clinical practice and further research in this area are explored.

METHODS OF THE REVIEW

Identification of Studies

A search was conducted of electronic databases, specialist journals and textbooks on myofascial pain syndrome. Forward citation tracking, reference scanning of articles included in the final sample, and contact with experts in the field, were used to identify additional articles. The search string for electronic databases included the following medical subject headings (MeSH) and text words (tw): (1) [Trigger Point (MeSH) OR trigger point (tw) OR myofascial pain (tw) OR muscle pain (tw)] AND [Physical Examination (MeSH) OR physical examination (tw) OR palpation (tw) OR manual examination (tw) OR Diagnosis, Musculoskeletal (MeSH) OR diagnosis (tw)]. To increase the sensitivity of the search, terms relating to the study design (eg, reliability) were excluded from the search string.

The OVID platform was used to search MEDLINE, EMBASE, the cumulative index of nursing and allied health literature (CINAHL), Alternative Medicine (AMED), and SPORTDiscus. A direct search of the Australasian Medical Index (AMI), and the Manual, Alternative and Natural Therapy Index (MANTIS) was also conducted. Each database was searched from inception until March 2008. The following journals and textbooks were manually searched for relevant articles:

The Journal of Musculoskeletal Pain (1998 to March 2008);
Journal of Musculoskeletal Medicine (1999 to March 2008);
International Musculoskeletal Medicine (1990 to March 2008);
Journal of Manual and Manipulative Therapy (1993 to March 2008);

Study Selection

Two reviewers (N.L., R.M.) independently screened all citations retrieved from the literature search, and identified those that satisfied the inclusion and exclusion criteria.

Inclusion Criteria

Studies that fulfilled all of the following criteria

1. investigated the reliability of physical examination (palpation) procedures designed to identify TPs.
2. used a repeated measures design and reported intraexaminer or interexaminer reliability.
3. investigated symptomatic or asymptomatic participants.
4. investigated “active” or “latent” TPs.

Exclusion Criteria

Studies were excluded for either of the following reasons

1. only reported the reliability of identifying tenderness (because the diagnostic criteria of TPs includes more than just tenderness).
2. used technology or instrumentation to identify TPs (eg, algometry or fine needle electromyography).

The 2 reviewers met and resolved discrepancies concerning the potential eligibility of studies. Full text articles of potentially eligible studies were obtained and independently evaluated according to the eligibility criteria. Those studies that met the eligibility criteria formed the final sample, and were independently assessed by the 2 reviewers according to predefined quality assessment criteria (Appendix).

Data Extraction and Quality Assessment

Two reviewers (N.L., R.M.) independently extracted data and assessed the quality of the studies. Discrepancies regarding study quality were discussed and resolved if necessary by consultation with a third reviewer (N.B.). Both reviewers had experience in the diagnosis and treatment of myofascial pain syndrome and the design and implementation of reliability studies. Each study in the final sample of eligible papers was assessed using a 12-item quality appraisal checklist designed to assess the quality and applicability of studies of diagnostic reliability (Appendix). The face validity of this checklist was established by consultation with methodology experts and by comparison with quality appraisal checklists used in other systematic reviews of diagnostic reliability.\textsuperscript{31–36} This checklist was also developed with reference to the statement for standards for reporting studies of diagnostic accuracy (STARD),\textsuperscript{37} and the quality assessment of diagnostic accuracy studies (QUADAS)\textsuperscript{38} appraisal tool. Studies were not given an overall numeric quality score; instead each item was
considered separately and answered “yes,” “no,” “unclear,” or “not applicable.”

**Estimates of Reliability**

Statistical estimates of reliability were extracted from the primary studies and reported for the purpose of summarizing the available evidence. Estimates of reliability were evaluated in the context of the quality of the study and its applicability to clinical practice. For categorical data, \( \kappa \) is routinely reported as a chance-corrected index of agreement, and is interpreted qualitatively in this report as: almost perfect (\( \kappa = 0.81-1.00 \)), substantial (\( \kappa = 0.61-0.80 \)), moderate (\( \kappa = 0.41-0.60 \)), fair (\( \kappa = 0.21-0.40 \)), slight (\( \kappa = 0.00-0.20 \)), or poor (\( \kappa < 0.00 \)).

**Grouping of Studies**

The eligible studies were divided into 2 main groups: those that included both symptomatic and asymptomatic participants, and those that included asymptomatic participants only. This distinction was made on the following grounds. Only active TPs produce symptoms and therefore can only be found in symptomatic participants. Only those studies that assessed the reliability of palpation to identify active TPs in symptomatic participants could be applicable to real-life clinical practice. Latent TPs may theoretically occur in both symptomatic and asymptomatic participants and it is therefore essential that examiners are able to reliably distinguish between active and latent TPs. Studies that include only asymptomatic participants can only be used to investigate the reliability of latent TP identification. As the literature describes latent trigger points as an entity that can occur in asymptomatic participants, these studies were retained and summarized separately for what they might show, in principle, about the diagnosis of active TPs in symptomatic patients.

**RESULTS**

**Literature Search**

A total of 2591 citations were identified from the literature search. Of these, 2563 were excluded after review of the title and abstract because they did not satisfy the eligibility criteria or were duplicates of the same study. Of the remaining 28 articles, 20 were excluded after full text review because they did not report reliability data, only reported muscle tenderness, did not assess TPs, or did not differentiate between TPs and tenderness. Nine studies, reported in 8 articles, were included in the final sample (Fig. 1). The study characteristics of the final sample are summarized in Table 1. No additional studies were identified for inclusion from the reference lists of the final sample, forward citation tracking, or from direct contact with experts in the field. A single discrepancy between reviewers about the inclusion of one paper was resolved by discussion. In addition, consensus between reviewers regarding the quality of included studies and their applicability to clinical practice was achieved after discussion with the third reviewer.

Five studies were conducted before the proposed inclusion of a palpable nodule in the diagnostic criteria, and so this feature was not evaluated in those studies. In the remaining 4 studies, however, only included this feature as a diagnostic criterion.

**Quality and Applicability Results**

None of the studies satisfied all of the relevant criteria for study quality and applicability (Table 2). None of the included studies provided data on the interexaminer reliability for the diagnosis of active TPs in symptomatic patients. One study came close to providing this information, but did not give separate reliability estimates for active and latent TPs.

Some studies were deficient in what might be considered technicalities. Six studies did not explicitly mention if observers were blinded to clinical information or cues that did not form part of the test under evaluation. Three of these studies did not mention if examiners were blinded to the findings of other examiners. Failure to adhere to blinding would bias the results in the direction of higher reported reliability. However, these studies were nevertheless given the benefit of doubt, and accepted on the grounds that explicit statements of these criteria of quality were simply not included in the published manuscript.

**Studies With Symptomatic and Asymptomatic Participants**

Five studies examined a mixture of symptomatic and asymptomatic participants and each reported the \( \kappa \) statistic as an estimate of reliability (Table 3). In each of these studies, agreement was based on the ability of examiners to agree upon the presence or absence of a physical sign in a particular muscle, but not on their ability to agree about the location of the sign within the muscle. In 2 of these studies, the examiners were expert in the field of myofascial pain and TP diagnosis, with extensive prestudy training for the specific purpose of increasing reliability. These examiners are not representative of those who would normally use the test in practice, and the results are unlikely to be normally used in practice.
<table>
<thead>
<tr>
<th>Author (y)</th>
<th>Participants</th>
<th>Raters</th>
<th>Study Characteristics</th>
<th>Diagnostic Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al-Shenqiti and Oldham (2005)</td>
<td>n = 58 diagnosed with rotator cuff tendonitis</td>
<td>Physiotherapist with 11 y experience</td>
<td>Intrarater reliability for rotator cuff muscles</td>
<td>LT, TB, JS, LTR, PR, RP</td>
</tr>
<tr>
<td>Bron et al (2007)</td>
<td>n = 40; 32 shoulder pain, 8 controls</td>
<td>Physiotherapist (n = 3) with 29, 28, 16 y experience</td>
<td>Interrater reliability for location of 3 trigger points in rotator cuff muscles</td>
<td>TB-N, LT, RP, LTR, JS</td>
</tr>
<tr>
<td>Gerwin et alA (1997)</td>
<td>n = 25; symptomatic/asymptomatic</td>
<td>Physiatrists (n = 2), neurologists (n = 2) all experts</td>
<td>Interrater reliability</td>
<td>LT, TB, LTR, PR, RP, TrP</td>
</tr>
<tr>
<td>Gerwin et alB (1997)</td>
<td>n = 10; 7 symptomatic, 3 asymptomatic</td>
<td>Phase A, had discussion session before data collection</td>
<td>For phase A, for 10 paired muscles</td>
<td></td>
</tr>
<tr>
<td>Hsieh et al (2000)</td>
<td>n = 52; 26 low back pain patients, 26 asymptomatic</td>
<td>Chiropractors (n = 2) with 3 and 5 y experience, and physiatrists (n = 2) with 3 and 6 y experience each received 3 × 2 h classroom lectures on MPS</td>
<td>Interrater reliability</td>
<td>TB, LTR, RP</td>
</tr>
<tr>
<td>Lew et al (1997)</td>
<td>n = 58; asymptomatic volunteers</td>
<td>Dual qualified physiotherapist and osteopath (n = 2) expert in trigger point diagnosis</td>
<td>Interrater reliability for the trapezius muscle</td>
<td>LT, TB, RP</td>
</tr>
<tr>
<td>McKenzie et al (1997)</td>
<td>n = 6; children with cerebral palsy</td>
<td>Clinicians (n = 3) experience and qualifications unclear</td>
<td>Interrater reliability study for trigger points in 16 muscles</td>
<td>TB-N, LT, RP</td>
</tr>
<tr>
<td>Njoo et al (1994)</td>
<td>n = 124; 61 low back pain patients, 63 controls</td>
<td>General practitioner (n = 1) with training at a rheumatology university hospital; final year medical students (n = 4) trained with the GP</td>
<td>Interrater reliability for the presence of TP in the QL and GM muscles in patients with and without LBP</td>
<td>LT, TB, JS, LTR, PR, RP</td>
</tr>
<tr>
<td>Sciotti et al (2001)</td>
<td>n = 20; asymptomatic volunteers</td>
<td>Clinicians (n = 4) (discipline or qualifications not specified), with 9, 13, 14, 18 y of experience Undertook 4 × 3 h training sessions on the testing procedure</td>
<td>Interrater reliability to identify the location of TP within the trapezius muscle</td>
<td>TB, LT, N, RP, LTR, JS</td>
</tr>
</tbody>
</table>

GM indicates gluteus maximus; GP, general practitioner; JS, jump sign; LT, local tenderness; LTR, local twitch response; MPS, myofascial pain syndrome; N, nodule; PR, pain reproduction; QL, quadriceps lumborum; RP, referred pain; TB, taut band; TP, trigger point.
to be applicable to everyday clinical practice. Nevertheless, these studies provide estimates of reliability that might constitute a benchmark of what could, or should, be obtained in conventional practice. Only 3 studies\textsuperscript{58,60,62} used what might be construed as representative examiners.

Estimates of reliability varied widely between each study, for the individual muscle tested, and also for each of the individual signs that make up the criteria for the identification of TPs. None of the studies stratified their results according to asymptomatic and symptomatic groups, nor active or latent TPs. Therefore, the extent to which estimates of reliability pertain to active or latent TPs is not evident. In the majority of studies, the outcome was dichotomous with the relevant feature being recorded as either absent or present, without specifying how indeterminate results were handled. In one study,\textsuperscript{58} all indeterminate results were pooled into the “absent” category, potentially inflating observed agreement and estimates of reliability between examiners.

Only one study reported agreement between examiners for the presence or absence of a TP within a group of muscles (\(k = 0.66-0.95\)), but not for the location of the TP.\textsuperscript{59} All other studies\textsuperscript{58,60,62} only reported estimates for the individual signs of TPs, without providing data for the overall agreement about the presence or absence of TPs. The small number of available studies differ markedly with respect to design, the prevalence of positive findings within the study sample, and the purported skill of the examiners, making comparisons between studies inappropriate.\textsuperscript{65} Hence, only a qualitative summary of the results is provided.

Across all studies, reliability estimates for the taut band ranged from \(k = 0.08\) to \(0.75\). For tenderness, the reliability estimates ranged from \(k = 0.22\) to 1.0, and for reproduction of familiar pain the reliability estimates ranged from \(k = 0.57\) to 1.0. For referred pain, reliability estimates ranged from \(k = 0.13\) to 0.84, and for the local twitch response estimates ranged from \(k = 0.05\) to 0.57. Reported estimates for the “jump sign” ranged between \(k = 0.07\) and 0.71.

Because \(k\) is affected by prevalence, we had aimed to report the frequency of abnormal findings in conjunction with \(k\) to aid interpretation. However, the studies did not provide sufficient data to compute the relevant frequencies.

Studies With Asymptomatic Participants Only

Two studies used asymptomatic volunteers.\textsuperscript{61,63} In one of these studies,\textsuperscript{64} the examiners identified latent TPs in the trapezius muscle and instructed an assistant to mark the location of the TPs on a printed body map that was reduced to approximately 25% of an average human adult. The accuracy and resolution of this recording method was not stated. In the other study,\textsuperscript{63} the examination was limited to the assessment of latent TPs in the right upper trapezius muscle. Once an examiner had identified a latent TP, an assistant applied a digitizing probe to the site indicated by the examiner and the 3-dimensional coordinates of the location were recorded electronically. In both of these studies,\textsuperscript{61,63} it was unclear if examiners were representative of those who would normally use the test in practice, and as all participants were asymptomatic, the findings of these studies lack applicability to everyday clinical practice. In addition, intramuscular fascial bands and neurovascular structures within the trapezius may account for the clinical
TABLE 3. Kappa Scores From Studies of Interrater Reliability

<table>
<thead>
<tr>
<th>Study</th>
<th>Quality Criteria (n = participants)</th>
<th>Tenderness</th>
<th>Pain Recognition</th>
<th>Referred Pain</th>
<th>Jump Sign</th>
<th>Taut Band</th>
<th>Local Twitch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bron et al (n = 40)</td>
<td>8/9</td>
<td>•</td>
<td>•</td>
<td>−0.13, 0.64</td>
<td>0.07, 0.68</td>
<td>0.11, 0.75(1)</td>
<td>−0.05, 0.45</td>
</tr>
<tr>
<td>Hsieh et al (n = 52)</td>
<td>7/9</td>
<td>•</td>
<td>•</td>
<td>0.13, 0.50(1)</td>
<td>•</td>
<td>0.00, 0.19(1)</td>
<td>0.02, 0.18(1)</td>
</tr>
<tr>
<td></td>
<td>7/9</td>
<td>•</td>
<td>•</td>
<td>0.08, 0.49(2)</td>
<td>•</td>
<td>−0.08, 0.06(2)</td>
<td>−0.02, 0.07(2)</td>
</tr>
<tr>
<td>Njoo et al (n = 124)</td>
<td>5/9</td>
<td>0.58, 0.73</td>
<td>0.57, 0.58</td>
<td>0.36, 0.46</td>
<td>0.68, 0.71</td>
<td>0.47, 0.51</td>
<td>−0.02, 0.19</td>
</tr>
<tr>
<td>Gerwin et alA</td>
<td>4/9</td>
<td>0.22*</td>
<td>†</td>
<td>†</td>
<td>†</td>
<td>†</td>
<td>†</td>
</tr>
<tr>
<td>(n = 25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gerwin et alB</td>
<td>4/9</td>
<td>0.48, 1.0</td>
<td>0.79, 1.0</td>
<td>0.57, 0.84</td>
<td>•</td>
<td>0.40, 0.46</td>
<td>0.11, 0.57</td>
</tr>
<tr>
<td>(n = 10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range across all studies</td>
<td></td>
<td>0.22, 1.0</td>
<td>0.57, 1.0</td>
<td>−0.13, 0.84</td>
<td>0.07, 0.71</td>
<td>−0.08, 0.75</td>
<td>−0.05, 0.57</td>
</tr>
</tbody>
</table>

*Indicates that this is the average κ obtained across 20 muscles. For Bron et al,*8 N indicates that examiners reported a palpable nodule within a taut band.
†Indicates that κ was not reported, but the authors reported that κ was poor. For Hsieh et al,*6 κ scores were provided for: (1) examiners agreement with an expert; and (2) agreement between examiners, excluding the expert. For Gerwin et al,*2 2 studies, A and B, were reported in 1 paper.

Studies are in descending order for total number of possible positive quality and applicability items achieved. For each sign, the range of κ scores that were obtained from the examination of each muscle included in the study is reported.

Criteria for which no κ scores were provided are indicated by •.

Neither of these two studies reported conventional measures of reliability. Lew et al*6* argue that κ could not be calculated from their data due to the presence of a “zero” cell in their contingency table, however this opinion is arguable. Our calculation from the data provided by Lew et al for the identification of latent TPs is κ = −0.26. Irrespective of this, Lew et al reported observed agreement (P0) of only 21% for latent TPs that produced referred pain, and 10% for latent TPs that did not produce referred pain. Importantly, in only 2 of 42 participants identified by both examiners as having latent TPs was there agreement about the actual location of the TPs. In the study of Sciotti et al,*5 agreements were assessed at 2 levels. First, examiner agreement was assessed for the presence or absence of latent TPs in the right upper trapezius muscle; however the data provided did not allow calculation of κ. Second, examiners were assessed on their ability to identify the location of a TP using a 3-dimensional computerized tracking system. The average of the 3-dimensional coordinates obtained from each examiner was used as a reference point to represent the estimated location of the TP in that participant. The distance by which each examiner deviated from the reference point was then used to calculate examiner agreement. The examiner error associated with identifying the estimated location of TPs ranged from 3.3 to 6.6 cm.

**Other Studies**

Al-Shenqiti and Oldham*57* reported the only study on intrarater reliability. Patients with a clinical diagnosis of rotator cuff tendinitis were evaluated for the signs of TPs within the rotator cuff muscles. Agreement was based on the identification of physical signs within a muscle, but not on the location of those signs within the muscle. The examiner was not blinded to the participants’ clinical history and it was not evident if the examiner was blinded to the identity of participants previously examined, or additional cues that may have influenced the test outcome. Information regarding the qualifications, training, skills and experience of the examiner were not adequately reported. The κ statistics reported for each sign ranged from κ = 0.75 to 1.0, and were higher than those from interrater reliability studies. However, it is difficult to interpret these results due to issues with the quality of this study, outlined above, and its applicability to clinical practice.

McKenzie et al*64* provided an estimate of examiner agreement for the presence or absence of TPs in children with cerebral palsy. The authors report that out of 16 muscles examined in 6 children by 3 examiners, only 5 disagreements for the presence or absence of TPs occurred, and reported an intraclass correlation coefficient of 0.93. However, examiner reliability was not the main focus of this paper and no information was included about blinding, or how the location of the physical signs were recorded and compared between examiners. In addition, there was inadequate detail to assess if appropriate statistical analysis was performed. Lastly, the participants included in this study were diagnosed with a neurologic disorder affecting muscle tissue and are not representative of patients in typical clinical practice. The results of this study are therefore difficult to interpret, and are reported here for the purpose of completeness.

**DISCUSSION**

None of the 9 studies in this review specifically reported intrarater reliability estimates for the identification of the location of active TPs in symptomatic participants. In those studies that did include symptomatic participants, the data were not stratified according to the symptom status of the participant, or the status of the TP as active or...
latent. Each of the included studies falls short of the optimal design in at least one respect thereby compromising the quality of the evidence.

Each study that examined both symptomatic and asymptomatic participants provided reliability estimates for individual signs of TPs. Only one of these studies provided estimates of examiner agreement about the presence or absence of a TP, which varied depending upon the muscle examined ($\kappa = 0.66-0.95$). However, the highest reported reliability estimate for the taut band in this study was $\kappa = 0.46$, which is lower than the range reported for the identification of a TP. It is therefore unclear which of the individual signs of TPs influenced the examiners in their overall decision about the presence or absence of a trigger point. The subjective signs derived from patient responses in this study, such as reports of tenderness ($\kappa = 0.48-1.0$), may have influenced examiners more than the objective signs derived from palpation, such as the taut band ($\kappa = 0.46$). Further, these data do not pertain to the currently proposed criteria for trigger points, which require the palpatory identification of a nodule within the taut band in addition to the reproduction of the patients' familiar pain; and nor do these data pertain to examiners agreement about the exact location of the TP. Lastly, the examiners in this study were expert in trigger point diagnosis and participated in prestudy training sessions for the purpose of increasing their reliability. For these reasons, the results from this study may not be applicable in everyday clinical practice and are likely to represent the upper limits of reliability estimates for TP diagnosis.

The remaining data pertain only to the reliability of individual signs of TPs, and not to the identification of TPs in totality. Overall, reliability was better for the subjective signs of tenderness, pain recognition, referred pain, and the jump sign, whereas reliability estimates for the taut band and twitch response were lower. Better reliability was recorded in studies that used experts who underwent special training than in studies that used conventional practitioners; and this underscores the concern that practitioners in typical clinical practice may fail to achieve the benchmarks set by experts.

In addition, and particularly prejudicial to TP theory, is the lack of any data on the reliability of pinpointing the exact location of active TPs. The relevant data on reliability pertain only to agreeing if a muscle has the signs of a TP, and not to the exact location of these signs. Yet independent identification of the location of the TP is essential to demonstrate that examiners are identifying the same phenomenon. If it is not evident that examiners can agree on the precise location of an active TP, they cannot be relied upon to place their treatment accurately. Two studies evaluated agreement between examiners as to the location of latent TPs. In one study, examiner agreement was found not to exceed 21%, and the other reported the overall magnitude of errors between examiners to be 3.3 cm to 6.6 cm, which is not sufficiently accurate for specific treatment such as injection.

**Implications**

Evidence for the diagnostic reliability of TPs is available from only a limited number of studies, and reliability data are lacking for clinically relevant, active TPs in symptomatic patients. In consideration of this, the diagnosis and treatment of TPs does not have a firm clinical basis. Until a reliable diagnostic test for TPs has been demonstrated it is recommended that this diagnosis should not be considered as a primary, or exclusive diagnosis for patients presenting a report of pain. If a treatment or management plan is to be implemented on the basis of a diagnosis of TPs, then patients should be informed of the ambiguity of this diagnosis so that they may make an informed choice about their treatment options.

In the absence of a reliable diagnostic test for TPs, one or more of 3 main problems will confound research investigations. First, they will be confounded by the inclusion of inappropriate cases. Second, it will not be possible to determine the precision of the diagnosis, or the treatment. Third, there will be no certainty in the measurement of TP resolution after treatment. This has implications for the interpretation of prior studies investigating the prevalence of TPs in specific patient populations, and those evaluating the effectiveness of treatments for TPs in clinical trials, for these studies are based on a test of unknown reliability and validity. Reliable methods of identifying TPs should be demonstrated before the implementation of further studies investigating the prevalence or treatment of trigger points.

For practical and taxonomic purposes, proponents of TP theory might choose to redefine the diagnostic criteria to include only tenderness and pain reproduction and eliminate the need for the taut band and twitch response. If the criteria are relaxed to define TPs as a tender area in muscle that when palpated reproduces the patients familiar pain, then the present evidence indicates that worthwhile agreement might be achieved. In that event, however, the TP theory would be compromised. Even though examiners agree upon tenderness and pain reproduction, these physical signs alone do not point to a pathologic entity, and would not be distinguishable from nonspecific muscle hyperalgesia.

**Future Research**

For a meaningful estimate of reliability for the diagnosis of TPs, an optimal study would have the following design features. Examiners would be blinded as to the condition of the patient, and to any clinical information other than that which they would derive from their examination. Without 37d8f82a9b837838, the study would not be a test of the clinical signs in question, for the other information might influence the examiners. Data would be collected and reported on the agreement, or otherwise, of all examiners for all participants. The raw data should be provided, so that consumers of the information are able to check and verify the estimates of reliability. For an entity whose diagnostic criteria are multiple and compound, data would be provided not only on the reliability of each component but also on the reliability of the composite findings simultaneously. For an entity such as TPs, agreement would be assessed for the identification of the exact location of the feature, for treatment is applied to the entity, and not generally to the muscle in which it is found. For an entity purported to produce symptoms, reliability would be tested in symptomatic participants, for it is only in such participants that the results apply and are of relevance. The examiners would be representative of the practitioners interested in the information, for only then might the results be applicable to everyday clinical practice. If experts are used, the results represent what the benchmark of performance might be, but they might not be
generalizable to less trained or less experienced practitioners.

The findings of this study prompt consideration of alternative explanations for the clinical phenomena attributed to trigger points. Other competing explanations include referred pain of peripheral nerve origin, secondary hyperalgesia, underlying joint tenderness, and normal intramuscular anatomy. A discussion about future research should therefore include reference to the validity of physical examination for TPs. Because there is no accepted reference standard for TPs, physical examination findings cannot yet be validated, even though they may be shown to be reliable. Researchers who investigate reliability should also consider comparing physical examination findings with the laboratory and imaging findings publicized as evidence of the existence of TPs.

CONCLUSIONS

No study to date has reported the reliability of TP diagnosis according to the currently proposed criteria in symptomatic patients. Although some of the individual signs, such as tenderness and pain reproduction, may be reliable to some extent in some muscles, their detection in isolation does not qualify as a diagnosis for TPs. The interrater reliability estimates for each sign varied widely for each muscle and between different muscles, ranging from less than chance to substantial reliability. However, insufficient data were provided upon which to interpret these results with confidence. On the basis of the limited number of studies available, and significant problems with their design, reporting, statistical integrity, and clinical applicability, physical examination cannot currently be recommended as a reliable test for the diagnosis of TPs. High-quality studies are needed to investigate the reliability of identifying the exact location of TP using current diagnostic criteria in clinically relevant patients.

Key Messages
- No study has reported the reliability of identifying the exact location of active TPs
- No study has reported the reliability of all diagnostic criteria simultaneously
- The majority of studies did not report important aspects such as whether examiners were blinded to clinical information or additional cues that may have influenced their findings
- For each individual criterion, estimates of reliability were inconsistent and ranged from excellent to less than chance for different muscles and across different studies
- Tenderness and pain reproduction achieved acceptable reliability; however these are not specific for TPs

APPENDIX

Quality Appraisal of Diagnostic Reliability Checklist

<table>
<thead>
<tr>
<th>Item</th>
<th>Yes</th>
<th>No</th>
<th>Unclear</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Was the test evaluated in a spectrum of subjects representative of patients who would normally receive the test in clinical practice?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>2. Was the test performed by examiners representative of those who would normally perform the test in practice?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>3. Were raters blinded to the reference standard for the target disorder being evaluated?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>4. Were raters blinded to the findings of other raters during the study?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>5. Were raters blinded to their own prior outcomes of the test under evaluation?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>6. Were raters blinded to clinical information that may have influenced the test outcome?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>7. Were raters blinded to additional cues, not intended to form part of the diagnostic test procedure?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>8. Was the order in which raters examined subjects varied?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>9. Were appropriate statistical measures of agreement used?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>10. Was the application and interpretation of the test appropriate?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>11. Was the time interval between measurements suitable in relation to the stability of the variable being measured?</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>12. If there were dropouts from the study, was this less than 20% of the sample.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
</tbody>
</table>

TOTAL

ACKNOWLEDGMENTS

The authors thank Ruth Mitchell, Cochrane Renal Group, The Children’s Hospital at Westmead, Sydney, Australia, and Phillipa Powis, Department of Clinical Research, Royal Newcastle Center and University of Newcastle, Newcastle, Australia, for their advice regarding the development of the search strategy.

REFERENCES


