Archived at the Flinders Academic Commons:
http://dspace.flinders.edu.au/dspace/

‘This is the peer reviewed version of the following article:
Carol Ann Flavell, Susan Gordon, Laurence Marshman & Kerrianne Watt (2014) Inter-rater reliability of classification systems in chronic low back pain populations, Physical Therapy Reviews, 19:3, 204-212,
DOI:10.1179/1743288X13Y.0000000131

which has been published in final form at
http://dx.doi.org/10.1179/1743288X13Y.0000000131

“This is an Accepted Manuscript of an article published by Taylor & Francis in Physical Therapy Reviews on 1 Jan 2014, available online: http://www.tandfonline.com/10.1179/1743288X13Y.0000000131.

© W. S. Maney & Son Ltd 2014
Systematic Review

Inter-rater reliability of classification systems in chronic low back pain populations

Carol Ann Flavell¹, Susan Gordon¹, Laurence Marshman²,³, Kerrianne Watt¹

¹School of Public Health Tropical Medicine and Rehabilitation Sciences, James Cook University, Townsville Campus, Douglas, Qld, Australia,
²Department of Neurosurgery, The Townsville Hospital, Angus Smith Drive, Townsville, Qld, Australia,
³School of Medicine, James Cook University, Townsville Campus, Douglas, Qld, Australia
ABSTRACT

Background: Low back pain (LBP) classification systems are used by physical therapists to classify patients. Classification systems require observation, and are at risk of rater bias and erroneous classification decisions, if the reliability among raters is poor. Rater reliability of individual systems in subgroups of LBP is important, to justify their continued utility.

Objectives: The purpose of this research was to investigate the reliability of LBP classification systems when applied exclusively to chronic low back pain (CLBP) populations.

Methods: A systematic electronic database search of Medline, CINAHL, PEDro, The Cochrane Library, Informit, and Scopus was conducted. Studies that reported reliability and detailed reliability statistics of one or more LBP classification systems, exclusively in CLBP populations were included. Two independent reviewers used the Quality Appraisal of Reliability Studies (QAREL) tool to evaluate quality and risk of bias for each study. Four eligible studies were identified.

Results: The Motor Control Impairment Classification System (OCS) and the Movement System Impairment Classification (MSI) were the only systems assessed for inter-rater reliability in CLBP populations. Inter-rater reliability for the MSI was substantial and inter-rater reliability for the OCS ranged from fair to almost perfect. However, risk of bias was high in the studies. Reported inter-rater reliability appeared to have an inverse relationship to study quality and risk of bias.

Conclusions: The findings of this review identified insufficient evidence to determine conclusions on interrater reliability when LBP classification systems are applied for CLBP. Therefore, recommendations to substantiate their use to classify patients reliably among therapists should be considered with caution.

Keywords: Back pain, Classification, Reliability, Review
INTRODUCTION

Classification systems for low back pain (LBP) have been described as structured clinical assessment pathways, which identify sub groups of patients.\textsuperscript{1} Physical therapists use LBP classification systems, which include, but are not limited to; evaluation of baseline symptom behavior, examination of movements, posture, neurological and neuro-dynamic testing and assessing spinal stability.\textsuperscript{2}

Several detailed classifications have been developed, with the belief that subgrouping people with LBP is important for both clinical and research purposes.\textsuperscript{3, 4} They include, but are not exclusive to, systems such as McKenzie’s mechanical diagnostic therapy (MDT).\textsuperscript{5} Sahrmann’s movement system impairment classification,\textsuperscript{6} O’Sullivan’s motor control impairment classification,\textsuperscript{7} classifications based on pain distribution,\textsuperscript{8} or patho-anatomical origin,\textsuperscript{9} and tools which classify LBP patients according to risk of chronicity,\textsuperscript{10} and predictors of outcome.\textsuperscript{11, 12}

Valid and reliable patient classification improves research methodology and constitutes best clinical practice by informing targeted interventions, which lead to improved patient outcomes. Predominantly, classifications used by physical therapists incorporate clinical examination techniques. Therapists implement many of these examination techniques as manual therapy or movement based treatment interventions, with response to treatment helping to predict patient outcome.

The process of subgrouping to homogenize research participants improves methodological rigor and ultimately research outcomes.\textsuperscript{3} LBP has often been grouped according to duration of symptoms. Three duration-based groups exist: acute, sub-acute and chronic low back pain (CLBP), yet further subgrouping within the three duration groups, based on movement patterns, and patho-anatomical origin, serves to homogenize LBP populations further.
Low back pain classification systems usually follow a detailed algorithm and guide treatment decisions via a clinical reasoning process, thereby reducing extraneous information gathering. Effective clinical reasoning is a key aspect of assessment; however, it is one with which undergraduate, newly graduated, and inexperienced health professionals struggle. Hence a defined yet flexible classification system will facilitate and provide clarity to the clinical reasoning process.

Classification systems that are reliable when applied to homogenous subgroups are imperative to achieve the best physical therapy practice and positive outcomes for patients, by way of more specific assessment and therefore focused interventions. This is particularly important for patients with CLBP, whose symptoms are of longer duration, and whose clinical presentation is often complicated by psychosocial factors, less often present in the acute or sub-acute. The prevention of chronicity in LBP has been a long established goal, yet difficult to attain. Physical therapists regularly encounter patients who’s LBP has already persisted to a chronic stage of greater than three months duration, and who have received little or no previous intervention by health professionals. These patients should receive the most appropriate methods of classification and treatment available. Hence, it is proposed that the reliability of existing classification systems in populations with CLBP should be established before recommendations for their use can be made. Accordingly, a study appraisal tool such as the Quality Appraisal of Reliability Studies (QAREL), which is designed specifically for reliability studies can be utilized to evaluate risk of bias for both internal and external validity, and statistical analysis.

This review was conducted to achieve two objectives. Firstly, to identify and appraise current literature which has evaluated the reliability of LBP classification systems when applied in homogenous populations of CLBP patients. Then subsequently, to identify the most reliable classification system for a proposed research project in a CLBP population.
METHOD

This systematic review was registered with the international prospective register of systematic reviews (CRD42013003655).

Search strategy

The following databases were accessed: Medline via OvidSP (1946 to September 2013), Cinahl (no date restriction), PEDro (no date restriction), the Cochrane library (no date restriction), Informit (1970 to September 2013) and Scopus (no date restriction). An initial keyword, title and abstract search using the search terms (truncated as required), low back pain, diagnosis, risk, classification, algorithm, develop, screening and reliability was conducted. A further search strategy was incorporated with the key terms, mechanical diagnosis and treatment, treatment based classification, patho-anatomical classification, movement system impairment classification, O’Sullivan classification system, and motor control impairment. Boolean operators ‘and’ and ‘or’ were applied between each search term (Table 1). The chief reviewer (CAF) conducted the database search on 9th September 2013.

Study selection

Potentially suitable articles were identified from their title and imported into Endnote (version 16). The chief reviewer (CAF) and second reviewer (SG) reviewed abstracts of identified articles. The reference lists of the identified abstracts were scanned for further suitable articles. Upon agreement full texts were sourced for inclusion by both reviewers. Disagreement amongst reviewers was resolved via consensus. Both reviewers appraised full text articles, for inclusion according to the following criteria.

Eligibility Criteria

Types of studies: Inter rater reliability studies of LBP classification systems that incorporated physical examination methods.
Participants: All study participants were required to be adults (≥ 18 years old), exclusively with LBP persistent for greater than 12 weeks.

Type of outcome measure or intervention: Studies that reported reliability and detailed reliability statistics for one or more LBP classification system were eligible for review.

Exclusion criteria

Systematic reviews, unpublished, non-peer reviewed publications, or opinion pieces, discussion papers, and studies not published in English were excluded. Exclusion applied to studies, which included any participants with symptoms of less than twelve weeks duration (acute and sub-acute LBP). Articles were also excluded for review, if participants were pre- or post-partum, diagnosed with inflammatory disease, malignancy or pain of visceral origin, or post-operative spinal surgery patients. Additional exclusions were any system based solely on clinical observation, self-reported questionnaires or other non-physical examination methods.

Data extraction process and review of methodological quality

Both reviewers extracted data from the eligible studies and appraised each using the QAREL data extraction form and checklist. The topics included in this appraisal form are participants, raters, blinding examination order, application and timing of tests, risk of bias, and use of appropriate statistics. Articles were scored as either “yes”, “no” or “unclear”. Some sections could be scored “not applicable”. “Yes” indicated good quality, “no” poor quality. The QAREL has been used in previous systematic reviews of reliability studies.

Both reviewers (CAF & SG) scored the final full text articles independently. Reviewers discussed and set the acceptable benchmarks for rating blinding and stability of variable sections on the QAREL checklist. Following independent review, any disagreement was resolved by consensus.
<table>
<thead>
<tr>
<th></th>
<th>Search strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Low back Pain* And Classification* And Diagno* And Reliability*</td>
</tr>
<tr>
<td>2</td>
<td>Low back Pain* And Classification* And Algorithm* And Reliability*</td>
</tr>
<tr>
<td>3</td>
<td>Low back Pain* And Classification* And Reliability* And Develop*</td>
</tr>
<tr>
<td>4</td>
<td>Low back Pain* And Algorithm* And Diagno* And Reliability*</td>
</tr>
<tr>
<td>5</td>
<td>Low back Pain* And Risk* And Screening*</td>
</tr>
<tr>
<td>6</td>
<td>Low back Pain* And Risk* And Reliability*</td>
</tr>
<tr>
<td>7</td>
<td>Low back Pain* And Screening* And Reliability*</td>
</tr>
<tr>
<td>8</td>
<td>Low back Pain* And Screening* And Reliability* And Algorithm*</td>
</tr>
<tr>
<td>9</td>
<td>Mechanical Diagnosis and Treatment* Or Treatment Based Classification* Or Patho-anatomical Classification* Or Movement System Impairment Classification* Or O'Sullivan Classification System* Or Motor Control Impairment</td>
</tr>
<tr>
<td>10</td>
<td>Low back Pain* And Mechanical Diagnosis and Treatment* Or Treatment Based Classification* Or Patho-anatomical Classification* Or Movement System Impairment Classification* Or O'Sullivan Classification System*</td>
</tr>
</tbody>
</table>
Synthesis and analysis of results

Quality Appraisal of Reliability Studies outcomes were summarized to allow comparison of study quality. In agreement with previous studies, 7 “yes” QAREL checklist items indicated moderate risk of bias. Consequently, < 7 “yes” items, indicated a high risk of bias, and ≥ 8 a low risk, and deemed to be of good quality.14, 17 The subsections of internal and external validity were scored separately and ‘yes’ scores calculated as a percentage of possible scores in that section. A section percentage of ≥ 67% defined the benchmark level for high quality, ≥ 50% for moderate, and < 50% poor quality in the studies.17, 18

RESULTS

Study Selection

Using the search syntax previously described (Table 1), 2384 research articles were identified from the electronic databases. Screening and review was conducted according to the standardized Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Figure 1).19 Subsequent to title screening, inclusions from reference list scanning, and removal of duplicate articles, 84 studies were identified for abstract screening. Both reviewers (CAF & SG) assessed the abstracts independently. Reviewer disagreement on seven abstracts was resolved by consensus agreement. Twenty-two abstracts were identified as suitable for full text review. Both reviewers independently scanned the full text versions and screened for eligibility.

Following full text review, 19 articles were rejected.6, 10, 11, 20-35 Reasons for rejection included studies which did not evaluate reliability (n=6),6, 11, 21, 31, 34, 35 did not evaluate CLBP exclusively (n=10),20, 22, 23, 25-28, 30, 32, 33 did not evaluate a clinical examination process for classification (n=1), 34 or classified as a prediction of risk for chronicity (n=2).10, 11 Several studies reported on reliability of heterogeneous LBP populations but the number of participants in each symptom duration subgroup was not specified, making it impossible to evaluate the results.
Total number of articles identified from database searches = 2384

Excluded: Duplicates & Title screen

Total number after duplicates removed and title screening = 64

Included: Identified articles from scanned reference lists = 20

Total number of abstracts screened = 84

Abstracts excluded:
- Not relevant/eligible = 14
- Not reliability studies = 34
- Not CLBP = 12
- Systematic reviews = 2

Total number of full text articles screened = 22

Full text articles excluded:
- Not exclusively CLBP = 10
- Not reliability studies = 6
- Did not evaluate a clinical examination system = 3

Total number of studies included for qualitative synthesis = 3

Figure. 1 Flow chart of study selection process
specific to CLBP (n=8).\textsuperscript{20,22-24,26,27,29,32} In addition, the definition of CLBP was sometimes ambiguous within studies (n=1).\textsuperscript{23} One further study specified numbers of participants according to symptom duration, but was rejected because it evaluated a questionnaire-based tool, which did not include a physical examination component.\textsuperscript{33} Neither of these studies aligned with the eligibility criteria definition.

The reviewers initially disagreed on the eligibility of three of the 19 excluded studies.\textsuperscript{21,23,24} The authors of the three contentious articles were contacted to clarify details of the classification process\textsuperscript{21}, or whether significant findings from the CLBP participants in the study population could be reported.\textsuperscript{23,24} Response from two of the three author groups was received. Subsequently, and in consideration of all available information, consensus was reached by the reviewers. Three full text articles of studies conducted exclusively on CLBP populations were accepted for final review. A description of the studies is provided in Table 2.

Two classification systems were evaluated in the studies accepted for review. Two studies reported on inter-rater reliability of the Movement System Impairment Classification (MSI),\textsuperscript{36,37} the third reported on the Motor Control Impairment Classification system (OCS).\textsuperscript{38} This study by Dankaerts et al.\textsuperscript{38} consisted of two separate reliability studies which were evaluated as ‘study 1’ and ‘study 2’.

\textbf{Risk of bias within studies}

The QAREL checklist of eleven items was applied to each of the four studies. Disagreement between the two reviewers occurred on five single items (11%). Disagreement was resolved by consensus discussion. Results of the QAREL evaluation for risk of bias ranged between six (low) and eight (high) (Table 3). Dankaerts et al.\textsuperscript{38} study 1, assessed the reliability of ‘expert’ raters (n=2) to classify CLBP using the OCS system. Dankaerts et al.\textsuperscript{38} study 2 in this article assessed the reliability of raters who were ‘moderately familiar’ and ‘very familiar’ with the OCS system (n=17). Study 1 showed a high risk of bias but study 2 a low
risk of bias. Both studies that used the MSI classification system, showed a high risk of bias.

Internal validity, external validity and statistical methods were evaluated using the QAREL. The external validity of all studies was considered high. However, internal validity varied between classifications, with the MSI classification studies by Trudelle-Jackson et al.,\textsuperscript{36} and Harris-Hayes and Van Dillen,\textsuperscript{37} showing internal validity of 60\%, compared to Dankaerts et al.\textsuperscript{38} study 1 and 2 of the OCS, which showed internal validity of 80\% and 100\% respectively. Statistical analysis rated highly (100\%) in both except Dankaerts et al.\textsuperscript{38} study 1.

**Summary of results**

All studies examined inter-rater reliability and Kappa values ranged from 0.32 to 0.96 (Table 2). Percentage of agreement was reported in all studies and ranged between 44\% and 97\%. The results of Dankaerts et al.\textsuperscript{38} study 1 conducted with two expert raters, showed almost perfect agreement ($k=0.96; 97\%$ agreement) but with high risk of bias.

In study 2 of the same article when ‘moderately familiar’ and ‘very familiar’ raters reviewed subjective case reports plus video examination, results showed, moderate ($k=0.55; 65\%$ agreement) and substantial ($k=0.71; 78\%$ agreement) agreement respectively. When the same raters reviewed subjective case reports only, fair agreement was reported for both ‘moderately familiar’ ($k=0.28; 44\%$) and ‘very familiar’ ($k=0.4; 54\%$ agreement) raters. Agreement for the MSI system was substantial in both studies, ($k=0.61$ and 75\% agreement; \textsuperscript{36} $k=0.75$; and 83\% agreement\textsuperscript{37}).
Table 2: Summary description and results of included articles

<table>
<thead>
<tr>
<th>Study</th>
<th>Objective</th>
<th>Classification system</th>
<th>Population</th>
<th>Raters</th>
<th>Method</th>
<th>Statistical analysis</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dankaerts et al.(^{38}) (study 1)</td>
<td>Determine inter-rater reliability of expert raters for classification of participants with NS-CLBP* with OCS</td>
<td>OCS</td>
<td>N = 35,18 women, All NS-CLBP*</td>
<td>Two musculo-skeletal physical therapists. One the developer of the system and 18 yrs experience in LBP. One extensive training by the developer and 12 yrs experience in LBP.</td>
<td>Raters were blinded. Re-examination 24hrs to 1 week.</td>
<td>Five diagnostic outcome variables (categorical) Kappa (k) &amp; percentage agreement</td>
<td>Agreement (%) = 97 k = 0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean duration 5.6 yrs, Age (y) 37 (12.73), ODI(^{†}) (%) 37 (11)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dankaerts et al.(^{38}) (study 2)</td>
<td>Determine inter-rater reliability of clinicians for classification of participants with NS-CLBP* with OCS</td>
<td></td>
<td>N = 25 Summary not given</td>
<td>N = 8 raters, 1 GP(^‡), 1 clinical neurologist, 3 musculo-skeletal physical therapists, 2 physical therapists. ‘Moderately familiar’ with CS. Training by clinical workshop with developer and instruction package N= 5 raters, 4 musculo-skeletal physical therapists, 1 sports physical</td>
<td>Video and case reports evaluated. Raters were blinded. Initially rated on case report only. Followed by combined case report and video examination.</td>
<td>Five diagnostic outcome variables (categorical) Kappa (k) &amp; percentage agreement between expert raters (gold standard) and other raters. Also agreement with both case report only or</td>
<td>1. All clinicians: a. Case report only Agreement (%) = 48 k = 0.32 b. Case report &amp; video Agreement (%) = 70 k = 0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2. Between clinician types: a. ‘Moderate familiar’ Case report only Agreement (%) = 44 k = 0.28</td>
</tr>
<tr>
<td>Trudelle-Jackson et al.</td>
<td>Determine inter-rater reliability for classification of participants with CLBP</td>
<td>MSI</td>
<td>N=24. 16 women. All CLBP (&gt; 12 weeks). Mean duration 288 wks. Age (y) 43.8 (13.5), ODI† (%) 37.4 (17.8),</td>
<td>Two physical therapists. Experience varied. Both trained in system use via courses. One rater trained by system developer. Raters practiced together on student subjects &gt;8hrs over 2 weeks prior to study.</td>
<td>Raters were blinded. 25 test items. Examinations conducted sequentially. No rest period and same day.</td>
<td>Five diagnostic outcome variables (categorical) Kappa (k) &amp; percentage agreement</td>
<td>Agreement (%) = 75 k = 0.61 (p&lt;. 0001) 95% CI 0.33-0.89</td>
</tr>
<tr>
<td>Harris-Hayes and Van Dillen</td>
<td>Determine inter-rater reliability for classification of</td>
<td>MSI</td>
<td>N=30. 21 women. No duration of symptoms reported. Stated as CLBP.</td>
<td>Two physical therapists, both with &gt;10yrs musculo-skeletal experience. One rater was the developer of the system, the other had</td>
<td>Raters were blinded. Examination on same day with 15 minute break between.</td>
<td>Five diagnostic outcome variables (categorical) Kappa (k) &amp; percentage agreement</td>
<td>Agreement (%) = 83 k = 0.75 (p&lt;.0001) 95% CI 0.51-0.99 z = 6.17</td>
</tr>
</tbody>
</table>
participants with LBP

<table>
<thead>
<tr>
<th></th>
<th>Mean Age (y)</th>
<th>ODI† (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>31.1 (12.9)</td>
<td>13.6 (7.5)</td>
</tr>
</tbody>
</table>

continuing education and 7 yrs experience using the system.

Training for the study was conducted by 2nd rater. Study of operations manual and practice with symptomatic and asymptomatic subjects.

CLBP: chronic low back pain

†NS-CLBP, †Oswestry Disability Index, †General Practitioner
### Table 3. Summary of risk of bias evaluation using Quality Appraisal of Reliability Studies (QAREL) checklist\textsuperscript{14}

<table>
<thead>
<tr>
<th>Diagnostic/classification approach</th>
<th>Dankaerts et al.\textsuperscript{38}</th>
<th>Study 1</th>
<th>Study 2</th>
<th>Trudelle-Jackson et al.\textsuperscript{36}</th>
<th>Harris-Hayes and Van Dillen\textsuperscript{37}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q 3-9: Internal Validity items:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Blinding:</strong></td>
<td>Interrater</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Intrarater</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>From reference standard</td>
<td>NA</td>
<td>Y</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>From other clinical information</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>From other cues</td>
<td>U</td>
<td>Y</td>
<td>U</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>Variation of examination order</td>
<td>Y</td>
<td>NA</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Suitable time interval between tests/examinations</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>U</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>4/5(=80%)</td>
<td>5/5(=100%)</td>
<td>3/5(=60%)</td>
<td>3/5(=60%)</td>
<td></td>
</tr>
<tr>
<td><strong>Q 1,2 &amp; 10: External Validity items:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suitable participant sample</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Suitable raters</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Appropriate test/examination conducted</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2/3(=66.66%)</td>
<td>2/3(=66.66%)</td>
<td>3/3(=100%)</td>
<td>2/3(=66.66%)</td>
<td></td>
</tr>
<tr>
<td><strong>Q 11: Statistics:</strong></td>
<td><strong>Total</strong></td>
<td>0/1(=0%)</td>
<td>1/1(=100%)</td>
<td>1/1(=100%)</td>
<td>1/1(=100%)</td>
</tr>
<tr>
<td><strong>Overall Total percentage “Yes”</strong></td>
<td><strong>Risk of Bias</strong></td>
<td>6</td>
<td>8</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Y: yes; N: no; U: unclear; NA: not applicable
DISCUSSION

This review identified limited evidence regarding the reliability of LBP classification systems when applied to homogenous populations with CLBP. Only reliability of the OCS and MSI classification systems were reported in the three eligible publications reviewed. A previous review which included reliability of low back pain classification systems, focused on non-specific low back pain. These authors provided detailed and important evidence on this topic. However, evaluation of study bias was not reported and the research was not exclusive to any particular duration based subgroup of non-specific LBP. Hence, to our knowledge no systematic reviews of LBP classification systems that have specifically evaluated, summarized and reported the risk of bias in studies of inter-rater reliability in CLBP populations exist. Accordingly, this review has relevance for clinicians and researchers, since it evaluated the reliability of LBP classifications systems when exclusively applied to CLBP populations. Without established reliability, a classification system cannot be recommended for either clinical use or research. In these circumstances, the assumed reliability of a LBP classification system in the absence of evidence cannot be guaranteed, which potentially affects outcomes of treatment interventions and research findings.

The two independent reviewers conducted a pre-appraisal discussion to reach agreement on the key aspects of the QAREL checklist as recommended by Lucas et al. The reviewers disagreed most often on item eight of the eleven-point list, ‘Was the order of examination varied?’ Four out of the five disagreements were related to this item. Criteria for item eight was clarified during the pre-appraisal discussion, but this item remained ambiguous. The authors are not aware of any research related to the reliability of the QAREL appraisal tool or any of its individual items despite its use in previous studies. Hence, the authors of this review support previous suggestions by Lucas et al. that further investigation be conducted to evaluate the reliability of the QAREL.
With the exception of Dankaerts et al. study 2, all articles showed a high overall risk of bias. The external validity was high for both the OCS and MSI systems. However, internal validity was lower for the MSI studies. This was related particularly to blinding from other cues not part of the classification, and the stability of the variable. In the studies by Trudelle-Jackson et al., Harris-Hayes and Van Dillen, and study 1 by Dankaerts et al., two raters conducted an examination of each participant. In study 2 by Dankaerts et al. rater classification was via pre-recorded reports and video, which resulted in the lower risk of bias. In support of the QAREL grading, the authors of this review believe that a re-examination process for classification provides a greater risk of bias from extraneous unintentional cues than do pre-recorded videos. Hence, the higher risk of bias grading for this section, due to the lack of clarity regarding avoidance or elimination of unintentional cues in all but study 2 by Dankaerts et al. In contrast, Clare et al. acknowledged that video or case reports may not absolutely reproduce all aspects of a face to face clinical examination. Nevertheless, this type of study provides consistent information, because it eliminates the risk of symptom aggravation, and the ability of participants to memorize previous symptom response between examinations.

Similar to the methods of Dankaerts et al. study 1, both of the MSI studies were conducted consecutively by two raters. However, unlike Dankaerts et al. the time intervals between examinations in the MSI studies were either unclear, or repeated fifteen minutes after the initial. A fifteen minute interval between examinations has been conducted in a previous study by Riddle and Rothstein, who evaluated inter-rater reliability of the MDT system. Nevertheless, in contrast to Trudelle-Jackson et al. it did not evaluate reliability specific to CLBP.

The high ‘irritability’ of acute LBP is often evidenced by its ease of symptom aggravation. This has been reported to lead to lower stability of examination variables due to rapid change in symptoms on movement testing. Despite the risk of bias from unintentional cues discussed
previously there is a convincing argument for shorter intervals between examinations in acute, sub-acute or even mixed cohort studies.\textsuperscript{20} In contrast, it is feasible to expect that the chronic nature of CLBP should present with lower ‘irritability’,\textsuperscript{42} and therefore a greater resultant stability of examination variables. Consequently, longer time intervals to reduce the risk of participant recall and attention bias,\textsuperscript{43} are indicated for studies with CLBP participants. The reviewers of this article suggest that it is theoretically and practically appropriate for reliability studies with CLBP participants, to conduct repeat examination at least one day, but no longer than two days following initial examination.

The results of this review indicate that where raters were either ‘experts’ or ‘very familiar’ with the classification, inter-rater reliability was substantial or almost perfect. These findings for CLBP are consistent with those of Fairbank et al.,\textsuperscript{44} who concluded that rater training contributed significantly to reported reliability in studies of LBP classification systems. This has important implications for clinical practice. It highlights the need for continued professional development and specific training in the utility of these systems. Thereby, clinicians will maximize reliability when applying classification systems, and subsequently target treatments for CLBP patients appropriately.

Variations in examination methodologies, re-examination time intervals, and experience and training in the application of the classification systems, limited the conclusions able to be drawn from this review. This is consistent with previous reliability studies, which investigated different LBP classifications in cohorts of mixed LBP duration.\textsuperscript{1, 20, 22-30, 32, 41, 45} Lack of consistency in current methodology limits comparison between studies and identifies a need to seek consensus on inter-rater reliability study design. This is particularly important given the paucity of studies specific to this LBP population.

Without doubt, the quest for prevention of chronicity in LBP is paramount. This has been supported recently by the work of\textsuperscript{10, 11, 33, 46} Nevertheless, circumstances prevail, whereby some
LBP patients first access physical therapy in a state of chronicity. Early and appropriate screening for risk and prevention of chronicity for all LBP patients is the ideal. However, in its absence there remains a need to improve the methods we currently implement to classify patients who present with CLBP. Limitations of this study include exclusion of articles not published in the English language. Nevertheless, this review established that minimal high quality evidence for inter-rater reliability of CLBP classification exists.

CONCLUSION

This review indicated that there is a lacuna of research reporting the inter-rater reliability of LBP classification systems when applied to CLBP populations. Currently the inter-rater reliability of only two systems, the OCS and MSI has been investigated in this population. The clinical implications of this review are that there is a lack of evidence for the reproducibility of these classifications for clinical use in this population. The implications for researchers are that the outcomes of randomized controlled trials that utilize these classifications systems to study CLBP populations will remain unsupported until risk of bias in studies is reduced and subsequently their reliability is established. Accordingly, to facilitate improved classification, management and outcomes for these patients, there is a critical need to conduct research on the reliability of existing LBP classification systems in CLBP populations.
DISCLAIMER STATEMENTS

Funding None

Conflict of interest No external financial support was provided. We know of no financial arrangements, organizational affiliations, or other relationships that might constitute a conflict of interest regarding the subject matter of the manuscript. We know of no special considerations regarding the submission, there are no related papers published or submitted for publication. We will not pursue publication of this article elsewhere until a final decision has been made by this journal.

Ethics approval None

Contributors Miss Carol Ann Flavell: Has made the greatest contribution to the concept and design of the review, and was one of the two critical reviewers. Conducted analysis and interpretation of review data. Drafted and revised the article. Approved the version to be published.

Associate Professor Gordon Susan: Contribution to the concept and design of the review, and was one of the two critical reviewers. Conducted analysis and interpretation of review data. Revised the article. Approved the version to be published.

Associate Professor Marshman Laurence: Contributed to the concept and design of the review. Conducted analysis and interpretation of review data. Revised the article. Approved the version to be published.

Associate Professor Watt Kerrianne: Contributed to the concept and design of the review. Revised the article. Approved the version to be published.
REFERENCES


