Reliability of procedures used in the physical examination of non-specific low back pain: a systematic review

May S, Littlewood C, Bishop A

CRD summary
The authors concluded that there were low levels of agreement between professionals about findings from most commonly used physical examination procedures used to assess patients with non-specific low-back pain. There were limitations to this review but, overall, the authors’ conclusions are likely to be reliable.

Authors' objectives
To determine the reliability of different physical examination procedures used to assess patients with non-specific low-back pain.

Searching
MEDLINE, PEDro, AMED, EMBASE, CINAHL and the Cochrane Library were searched from inception to August 2005 for studies reported in English; the search terms were reported. In addition, reference lists of identified studies were screened. Only studies published as full reports before August 2005 were eligible; abstracts were excluded.

Study selection

Study designs of evaluations included in the review
Inclusion criteria were not specified with respect to study design.

Specific interventions included in the review
Studies that evaluated physical examination procedures of the lumbar spine used for clinical decision-making were eligible for inclusion. Studies evaluating tests for specific back pain or instrumented procedures (other than timed endurance) were excluded. The included studies evaluated:

a variety of different tests of palpation (identifying spinal level, passive accessory movements, comparable level, passive physiological movements, muscle tension, muscle spasm, 'manipulative lesion'/flexion, misalignment, subluxation and instability tests of posterior shear or prone stability),

pain response (response to repeated movements, centralisation, directional preference, relevance of lateral shift/component, pain on movement or palpation, osseous pain, soft tissue pain and trigger point symptoms),

observation (timed muscle endurance in abdominals or back extensors, lateral shift/direction, lordosis, visual observation of abnormality, disturbance of normal lumbopelvic rhythm and coupling patterns), and

classification systems (McKenzie syndromes/subsyndromes, movement system impairment categories, treatment-based classification systems, Canadian Back Institute system and diagnostic classification system).

Some studies evaluated more than one examination procedure. Most of the physical procedures were conducted by physiotherapists; others were conducted by chiropractors, physicians, osteopaths or other professionals.

Participants included in the review
Studies of adults (aged older than 18 years) with non-specific back pain were eligible for inclusion in the review.

Outcomes assessed in the review
Studies that assessed intra- or inter-examiner reliability were eligible for inclusion. The review assessed reliability using various statistics: kappa, weighted kappa, intra-class correlation coefficient (ICC), Pearson's correlation coefficient (PCC), and Scott's P (SP). Most of the included studies reported kappa or ICC statistics.
How were decisions on the relevance of primary studies made?
One reviewer selected studies and a second reviewer confirmed the eligibility of those selected.

Assessment of study quality
Two reviewers independently assessed validity using criteria modified from those described by van der Wurff et al. and Bogduk (see Other Publications of Related Interest nos.1-2). Differences between reviewers were resolved by discussion or through recourse to a third reviewer. The criteria assessed were related to:

model/patient population (description, representative sample, random or consecutive selection and sample size),

test procedure (described and reproducible, uniform execution, measures to reduce bias and highest level of examiner, consensus procedure prior to testing with pilot study), and

test results (more than one pair of examiners tested, multiple testing between examiners, standardised measure of test outcome, reporting of frequencies of outcome and agreement, and use of appropriate interferential statistics).

The maximum quality score was 100 points; studies scoring 60 or more were considered to be higher quality. The criteria were pilot tested on 3 studies before being applied to all other studies.

Data extraction
The authors did not state how the data were extracted for the review, or how many reviewers performed the data extraction.

For each study, the reliability statistic or range of values for the statistic (kappa, weighted kappa, ICC, PCC or SP) was presented with 95% confidence intervals (CIs), where possible, for each physical examination reported. Some statistics were calculated from original data.

Methods of synthesis
How were the studies combined?
The studies were grouped by type of physical examination test and combined in a narrative. A cut-off value of 0.85 for kappa and the ICC was determined a priori to indicate satisfactory reliability. A sensitivity analysis was conducted by lowering the cut-off value for satisfactory reliability to 0.7 for kappa and the ICC. The authors appeared to grade the level of evidence for each intervention using a hierarchy of evidence: ‘strong’ reflected consistent findings from three or more high-quality trials; ‘moderate’ reflected consistent findings from low-quality trials and/or one high-quality trial; ‘limited’ reflected evidence from one low-quality trial; ‘conflicting’ reflected inconsistent findings among multiple trials; and no evidence reflected an absence of trials.

How were differences between studies investigated?
Differences between the studies were not formally investigated.

Results of the review
Forty-eight studies were included. Forty-seven studies assessed inter-tester reliability and 10 studies evaluated intra-tester reliability. The number of patients or testers was not reported.

The mean quality score was 52 (range: 0 to 88). Common flaws were sample not representing clinical practice, sample not selected randomly or consecutively, lack of standardisation of procedures and lack of control of bias. Forty per cent of the studies were classified as higher quality.

Taking the cut-off value for satisfactory reliability as 0.85, most procedures showed either conflicting evidence or moderate to strong evidence of low reliability.

Taking the cut-off value for satisfactory reliability as 0.70, there was moderate evidence (changed from contradictory)
about pain response to repeated movement.

**Authors' conclusions**
The most commonly used physical examination procedures for assessing patients with low-back pain had poor reliability.

**CRD commentary**
The review question was clearly defined. Several relevant sources were searched but no attempts were made to reduce publication or language bias. Methods were used to minimise reviewer errors and bias in the selection of studies and assessment of validity, but it was unclear whether similar steps were taken in the extraction of data. Study validity was assessed and the results reported. The authors stated that the patients were not representative of clinical practice, but no information was given about patients to illustrate this. In addition, the numbers of patients and testers were not reported and this hinders an independent evaluation of the evidence.

In view of the diversity of the studies, a narrative synthesis that took study quality into account was appropriate. However, a more detailed summary of the results would have been helpful rather than summing up the evidence as strong, moderate or weak. Where results differed among studies, potential reasons for these differences were not discussed. There were limitations to this review but, overall, the authors’ conclusions are likely to be reliable.

**Implications of the review for practice and research**

**Practice**: The authors did not state any implications for practice.

**Research**: The authors stated that the quality of studies assessing reliability could be improved by enrolling a large (>50) representative sample of consecutive or randomly selected patients, using clearly defined procedures, taking measures to reduce bias, using multiple pairs of testers, standardising test outcome measures, reporting frequencies and agreement, using appropriate statistics and adequately reporting the study.

**Bibliographic details**

**PubMedID**
16764546

**Other publications of related interest**

**Indexing Status**
Subject indexing assigned by NLM

**MeSH**
Humans; Low Back Pain /classification /diagnosis; Pain Measurement /methods; Palpation /methods; Physical Examination /methods; Physical Therapy Modalities; Reproducibility of Results

**AccessionNumber**
12006003741

**Date bibliographic record published**
30/11/2007

**Date abstract record published**
30/11/2007

**Record Status**
This is a critical abstract of a systematic review that meets the criteria for inclusion on DARE. Each critical abstract contains a brief summary of the review methods, results and conclusions followed by a detailed critical assessment on the reliability of the review and the conclusions drawn.