BJSM Online First, published on December 19, 2016 as 10.1136/bjsports-2016-096558

Reliability of specific physical examination tests for the diagnosis of shoulder pathologies: a systematic review and meta-analysis

Toni Lange,[1] Omer Matthijs,[2,3] Nitin B Jain,[4] Jochen Schmitt,[1] Jörg Lützner,[5] Christian Kopkow[1,6]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bjsports-2016-096558).

For numbered affiliations see end of article.

Correspondence to
Professor Christian Kopkow, Department für Angewandte Gesundheitswissenschaften, Hochschule für Gesundheit (University of Applied Sciences), Gesundheitscampus 6–8, Bochum 44801, Germany; Christian.Kopkow@hs-gesundheit.de

Accepted 28 November 2016

ABSTRACT

Background Shoulder pain in the general population is common and to identify the aetiology of shoulder pain, history, motion and muscle testing, and physical examination tests are usually performed.

Objective The aim of this systematic review was to summarise and evaluate intrarater and inter-rater reliability of physical examination tests in the diagnosis of shoulder pathologies.

Methods A comprehensive systematic literature search was conducted using MEDLINE, EMBASE, Allied and Complementary Medicine Database (AMED) and Physiotherapy Evidence Database (PEDro) through 20 March 2015. Methodological quality was assessed using the Quality Appraisal of Reliability Studies (QAREL) tool by 2 independent reviewers.

Results The search strategy revealed 3259 articles, of which 18 finally met the inclusion criteria. These studies evaluated the reliability of 62 test and test variations used for the specific physical examination tests for the diagnosis of shoulder pathologies. Methodological quality ranged from 2 to 7 positive criteria of the 11 items of the QAREL tool.

Conclusions This review identified a lack of high-quality studies evaluating inter-rater as well as intrarater reliability of specific physical examination tests for the diagnosis of shoulder pathologies. In addition, reliability measures differed between included studies hindering proper cross-study comparisons.

Trial registration number PROSPERO CRD42014009018.

BACKGROUND

Shoulder pain in the general population is common, with a reported prevalence of 7–26%.[1] Patients suffering from shoulder pain often are limited in performing activities of daily living and therefore seek help from healthcare professionals, resulting in substantial utilisation of healthcare resources.[2 3] To identify the aetiology of shoulder pain, history, motion and muscle testing, and physical examination tests are usually performed.[4] Physical examination tests aim to reproduce the patients symptoms (pain), which contrasts to other physical examination tests and outcome tests performed by clinicians, such as range of motion and muscle tests, as reviewed by Roy and Esculier.[5] Several systematic reviews have evaluated the validity of physical examination tests, concluding that most research is of insufficient methodological quality or that lacks consistently solid measures for validity obtained from studies with higher methodological quality.[6–11] Sciascia *et al*[12] performed a survey among orthopaedic shoulder surgeons and identified that a wide variety of tests were used to evaluate patients with shoulder symptoms.[12] Notably, lacking evidence regarding the diagnostic accuracy did not preclude use of the tests in clinical practice.[12] However, both validity, and reliability is of concern if physical examination tests are applied.[13 14] A poor reliability has a negative influence on the test's validity,[15] thus a test will not be valid if it does not measure consistently.[16] Tests with insufficient reliability (eg, training of examiners, variation in test execution due to examiners) might be the reason for varying results regarding the validity of physical tests.[17] To date, one systematic review published in 2010 evaluated the reliability of physical examination tests for the shoulder,[18] concluding that there is no consisting evidence that any tests have acceptable levels of reliability. Within the past few years more research on the reliability of physical examination tests has been performed and recent studies have been published. Therefore the objective of this systematic review is to systematically summarise and critically appraise research on the reliability of physical examination tests used for the diagnosis of shoulder pathologies.

METHODS

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were used.[19] The PRISMA statement aims to improve the reporting of systematic reviews and meta-analyses. This systematic review was registered a priori within the International Prospective Register of Systematic Reviews (PROSPERO; CRD42014009018).

Inclusion/exclusion criteria

Studies assessing the intrarater and/or inter-rater reliability of specific physical examination tests for the diagnosis of shoulder pathologies applied as a single test or in combination with other tests were included if written in English or German. Studies on patients of every age and setting were considered eligible. Studies were excluded if they did not name or describe the physical tests or did not refer a source that did so. Studies were excluded if the overall reliability of a group of tests was reported but individual tests were not specified/named or if the authors made use of generic terms such as physical examination to describe an unspecified combination of physical tests. In addition, studies

To cite: Lange T, Matthijs O, Jain NB, *et al*. *Br J Sports Med* Published Online First: [*please include* Day Month Year] doi:10.1136/bjsports-2016-096558

were excluded if only asymptomatic patients were evaluated or if the physical examination test was performed under anaesthesia or immediately postoperative. Animal studies and cadaveric studies and studies which used device supported testing procedures (defined as devices which are deemed too expensive or time-consuming for daily clinical practice) were also excluded.

Search strategy

A comprehensive systematic literature search in the following databases via the Ovid interface from inception until 18 March 2014 was performed, accessed via the Saxon State and University Library Dresden (SLUB): MEDLINE from 1946, EMBASE from 1974, and the Allied and Complementary Medicine Database (AMED) from 1985. The search strategy included terms about diagnostic tests, the conditions of interest, structures at risk, and reliability (see online supplementary appendix). Additionally, Physiotherapy Evidence Database (PEDro) was searched with a modified search strategy using the body part filter (upper arm, shoulder or shoulder girdle) in combination with the terms for reliability as used for the search in MEDLINE, EMBASE and AMED. Furthermore, reference lists of all eligible articles were screened for further relevant studies. A search update using the same search strategy and electronic databases was conducted on 20 March 2015 to identify recently published articles. The original search strategy was designed to identify studies on the reliability of specific physical examination tests evaluating specific structures (eg, rotator cuff tear) and general physical examinations tests (eg, strength or range of motion testing for the shoulder as well as shoulder girdle). However, in this review only the results of physical examination tests for the diagnosis of shoulder pathologies are reported.

Study selection and data abstraction

Identified titles and abstracts were screened by two independent reviewers (TL and CK), according to the described inclusion criteria. Subsequently, full texts were checked independently for eligibility by the same two reviewers. Any disagreements were resolved by consensus and if needed by a third reviewer (JS). Before titles and abstracts screening initiation, two subsamples consisting of randomly selected 50 titles and abstracts from all identified articles were performed. Afterwards the two reviewers (TL and CK) discussed their procedure to avoid following inequalities and started with the titles and abstract screening after an almost perfect agreement (according to classification system proposed by Landis and Koch[20]) was reached in the second pretest subsample (subsample 1: Cohen's κ=0.22, percentage agreement=88.00%; subsample 2: Cohen's κ=1.00, percentage agreement=100.00%). Data extraction was done by one reviewer (TL) and checked in duplicate by the other reviewer (CK). For standardised data extraction, forms were used, which were created according to the Quality Appraisal of Reliability Studies (QAREL) checklist.[21] Data on the objectives, patients, raters, physical examination tests, outcome variables and results were extracted. Authors of primary studies were contacted if additional data was needed. In case the authors provided the requested information, the appropriate reliability measures were calculated if possible.

Quality assessment

Quality assessment of all included studies was carried out independently by the two reviewers (TL and CK) using the QAREL checklist.[21] QAREL is especially designed for the quality assessment of reliability studies and is considered to be reliable for use.[22] The checklist consists of 11 items evaluating seven methodological domains of reliability studies (spectrum of patients and of examiners, examiner blinding, time interval between repeated measures, test application and interpretation, order of examination and statistical analysis of the data). Items can be answered with 'yes', 'no' or 'unclear' (and in addition if necessary with 'not applicable'). Fulfilled quality aspects of studies are indicated with a 'yes', whereas not fulfilled aspects with a 'no'. If insufficient information is provided to properly judge the quality aspect of studies, this is indicated with an 'unclear'. As recommended,[22] criteria by which judgments were made for each item of QAREL were a priori defined and tested by the two reviewers (TL and CK) (see online supplementary appendix).

If both intrarater and inter-rater reliability of a physical examination test was evaluated in one single publication, the quality assessment using QAREL was performed separately for (a) the intrarater and (b) the inter-rater reliability to account for specific possibilities for bias.

Data synthesis

Reliability measures are presented as reported by the authors of primary studies. Cohen's κ values <0.00 indicate poor, 0.00–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial and >0.81 almost perfect agreement.[20] Intraclass correlation coefficient (ICC) values <0.40 represent poor, values between 0.40 and 0.75 represent fair to good, and values above 0.75 represent excellent reliability.[23]

The agreement among reviewers of title and abstract screening was measured with percentage agreement and Cohen's κ statistic (95% CI), prevalence-adjusted bias-adjusted κ (PABAK), positive and negative percentage agreement as well as bias- and prevalence index. The agreement among reviewers of methodological quality using the QAREL tool was measured with percentage agreement and Cohen's κ statistic (95% CI).

Meta-analysis of Cohen's κ was performed according to the statistical framework proposed by Sun,[24] if raters in studies considered eligible for meta-analysis were clearly blinded to other raters.

All statistical analyses were performed using R V.3.2.0 (The R Project for Statistical Computing, Vienna, Austria) and RStudio (RStudio, Boston, Massachusetts, USA).

RESULTS

Study selection

Results of the literature search and study selection are shown in the PRISMA flow chart (figure 1). Agreement among reviewers regarding screening of titles and abstracts yielded a Cohen's κ of 0.76 (CI 0.70 to 0.82) and of full-text articles 0.68 (CI 0.54 to 0.81); additional reliability statistics are presented in figure 1. After full texts were reviewed, 18 publications met our criteria. These 18 studies presented data on 62 different physical examination tests and test modifications. Characteristics of included studies are summarised in online supplementary table S1.

All included studies were prospective and inter-rater reliability was assessed in 17,[25–41] and 1 study evaluated inter-rater and intrarater reliability.[42] In 16 of the 18 publications, primary care settings were used.[25–27 29–36 38–42] One study was performed in a tertiary care setting[37] and one in undefined care settings.[28]

Methodological quality

Results of methodological assessment using QAREL for all included studies are summarised in table 1. Methodological quality ranged from 2/11 rating[29] to 7/11 total positive ratings[30]

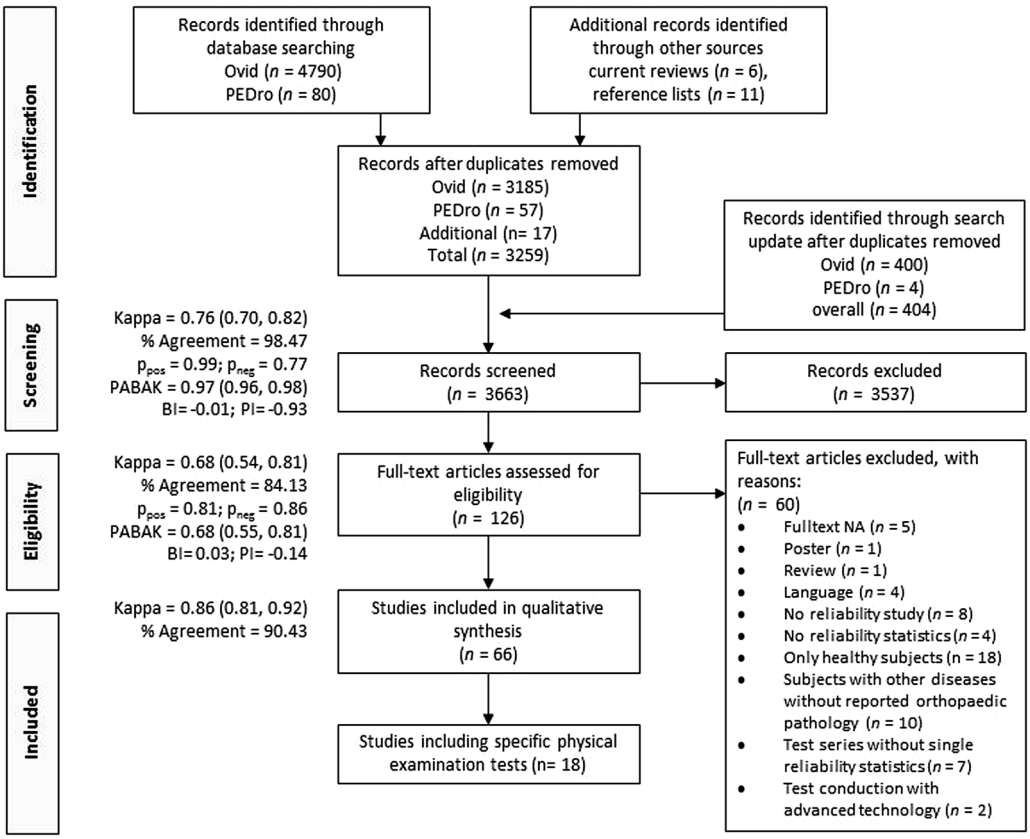

Figure 1 PRISMA flow chart. BI, bias index; NA, not applicable; neg, negative; PABAK, prevalence-adjusted bias-adjusted κ; PEDro, Physiotherapy Evidence Database; PI, prevalence index; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; pos, positive.

Table 1 Overview of risk of bias assessment used with QAREL checklist

		QAREL items											
Study	**Design**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Total**
Cadogan *et al*[25]	Inter-rater	Y	Y	Y	NA	U	U	U	Y	U	Y	Y	6
Dromerick *et al*[26]	Inter-rater	Y	Y	Y	NA	U	U	U	U	U	Y	Y	5
Johansson and Ivarson[42]	Inter-rater	Y	Y	U	U	U	U	U	U	Y	Y	Y	5
Johansson and Ivarson[42]	Intrarater	Y	Y	U	U	U	U	U	U	Y	Y	Y	5
Kim *et al*[27]	Inter-rater	Y	Y	Y	NA	Y	U	U	U	U	Y	Y	6
Kim *et al*[28]	Inter-rater	Y	U	Y	NA	Y	Y	U	U	U	Y	Y	6
Kim *et al*[29]	Inter-rater	Y	U	U	NA	U	U	U	U	U	Y	U	2
Kim *et al*[30]	Inter-rater	Y	Y	Y	NA	Y	Y	U	U	U	Y	Y	7
Michener *et al*[31]	Inter-rater	Y	Y	Y	NA	Y	N	U	U	U	Y	Y	6
Nanda *et al*[32]	Inter-rater	Y	Y	Y	NA	U	U	U	U	U	Y	Y	5
Nomden *et al*[33]	Inter-rater	Y	Y	Y	NA	U	U	U	U	U	Y	Y	5
Norregaard *et al*[34]	Inter-rater	Y	Y	U	NA	Y	U	U	Y	U	Y	Y	6
Ostor *et al*[35]	Inter-rater	Y	Y	U	NA	U	N	U	U	U	Y	Y	4
Palmer *et al*[36]	Inter-rater	Y	N	U	NA	U	U	U	U	U	Y	Y	3
Spencer[37]	Inter-rater	Y	Y	Y	NA	U	N	U	N	Y	Y	Y	6
Tzannes *et al*[38]	Inter-rater	Y	Y	Y	NA	U	U	U	U	U	Y	N	4
Vind *et al*[39]	Inter-rater	Y	U	Y	NA	U	U	U	N	U	Y	Y	4
Walker-Bone *et al*[40]	Inter-rater	Y	N	Y	NA	U	U	U	Y	U	Y	Y	5
Walsworth *et al*[41]	Inter-rater	Y	Y	Y	NA	Y	N	U	U	U	Y	Y	6

QAREL items: 1. Was the test evaluated in a sample of patients who were representative of those to whom the authors intended the results to be applied? 2. Was the test performed by raters who were representative of those to whom the authors intended the results to be applied? 3. Were raters blinded to the findings of other raters during the study? 4. Were raters blinded to their own prior findings of the test under evaluation? 5. Were raters blinded to the results of the reference standard for the target disorder (or variable) being evaluated? 6. Were raters blinded to clinical information that was not intended to be provided as part of the testing procedure or study design? 7. Were raters blinded to additional cues that were not part of the test? 8. Was the order of examination varied? 9. Was the time interval between repeated measurements compatible with the stability (or theoretical stability) of the variable being measured? 10. Was the test applied correctly and interpreted appropriately? 11. Were appropriate statistical measures of agreement used?
N, no; NA, not applicable; QAREL, Quality Appraisal of Reliability Studies; U, unclear; Y, yes.

and from 2/11 ratings[37] to 8/11 total unclear ratings.[29] Recruitment of raters was not specified in any of the included studies. In nine included studies, patients were recruited consecutively[25 26 29–34 37] and in one study through convenience sampling.[42] In three studies patients were referred[35 36 38] and in five studies the recruitment protocol was unclear.[27 28 39–41] Blinding of raters to the findings of other raters was unclear in 4 of the 18 inter-rater reliability studies.[29 34 35 42] In the intrarater reliability studies, the blinding of raters to their own prior findings was judged as unclear due to insufficient information.[42] Blinding to further clinical information was stated in 3 of the 18 included studies.[28 30 35]

Percentage agreement among reviewers regarding the rating of methodological quality of included studies using QAREL ranged for the different QAREL items from 74% to 100%. The overall agreement between raters of the methodological assessment using QAREL yielded a Cohen's κ of 0.86 (CI 0.81 to 0.92).

Physical examination tests

Physical examination tests for the diagnosis of shoulder pathologies were categorised as follows: acromioclavicular dysfunction tests, impingement tests, torn labrum/instability tests, and torn rotator cuff/impingement tests. Altogether 62 different physical examination tests were evaluated in the studies included in this systematic review (see online supplementary table S2 and table 2). Since only one study evaluated the intrarater reliability of physical examination tests for the diagnosis of shoulder pathologies (table 2),[42] comparisons between intra- and inter reliability was not possible. Cohen's κ was the most used reliability measure and was used in 77% of studies with categorical outcomes. Strength of agreement of acromioclavicular dysfunction tests ranged from slight to moderate agreement, impingement tests ranged from slight to almost perfect, torn labrum/instability tests ranged from poor to almost perfect and torn rotator cuff/impingement ranged from fair to almost perfect (see online supplementary table S2).

Meta-analysis identified extensive heterogeneity for the Hawkins-Kennedy Test, Neer Test, Empty Can Test/Supraspinatus Test, Painful Arc Test (figures 2–5) with I^2 values >0.75, which can be interpreted as 'considerable heterogeneity' according to the Cochrane Handbook.[43] Results from meta-analysis indicate moderate-to-substantial inter-rater reliability for the Hawkins-Kennedy Test, Neer Test, Empty Can Test/Supraspinatus Test and the Painful Arc Test.

DISCUSSION

Main findings

This systematic review identified 18 articles, which examined the reliability of 62 physical examination tests for the diagnosis of shoulder pathologies with varying inter-rater reliability. Intrarater reliability was investigated in only one study assessing four different tests, reporting almost perfect reliability. The included studies were of low methodological quality according to the QAREL tool.[21] Meta-analysis identified extensive heterogeneity among studies for physical examination tests using the I^2 statistic,[44 45] thus the findings of the meta-analysis may be inaccurate and need to be interpreted with caution. Results from meta-analysis indicate moderate-to-substantial inter-rater reliability for the Hawkins-Kennedy Test, Neer Test, Empty Can Test/Supraspinatus Test and the Painful Arc Test. These examination procedures (and other tests evaluated in this systematic review) need to be used with great caution in terms of

Table 2 Results of intrarater reliability

Index test	Study	Instruments	Test position	Prevalence	$N_{subject}$	$N_{shoulder}$	N_{rater}	N_{trial}	Estimates of reliability (95% CI)	P_o (%)	Measures of uncertainty	Strength of agreement	QAREL score
Impingement tests													
Hawkins-Kennedy Test	Johansson and Ivarson[42]	NA	Sitting	1.00	33	33	2	2	κ=1.00 (1.00 to 1.00)*; PABAK=1.00 (1.00 to 1.00)*	100	BI=0.00*; PI=0.52*	Almost perfectt	5/11
Neer Test	Johansson and Ivarson[42]	NA	Sitting	1.00	33	33	2	2	κ=1.00 (1.00 to 1.00)*; PABAK=1.00 (1.00 to 1.00)*	100	BI=0.00*; PI=0.58*	Almost perfectt	5/11
Patte Test	Johansson and Ivarson[42]	NA	Sitting	1.00	33	33	2	2	κ=1.00 (1.00 to 1.00)*; PABAK=1.00 (1.00 to 1.00)*	100	BI=0.00*; PI=0.88*	Almost perfectt	5/11
Torn rotator cuff/impingement													
Empty Can Test/ Supraspinatus Test	Johansson and Ivarson[42]	NA	Sitting	1.00	33	33	2	2	κ=1.00 (1.00 to 1.00)*; PABAK=1.00 (1.00 to 1.00)*	100	BI=0.00*; PI=−0.09*	Almost perfectt	5/11

*Calculated based on data obtained from author of primary studies.
†Landis and Koch.[20]
BI, bias index; NA, not applicable; PABAK, prevalence-adjusted bias-adjusted κ; PI, prevalence index; QAREL, Quality Appraisal of Reliability Studies.

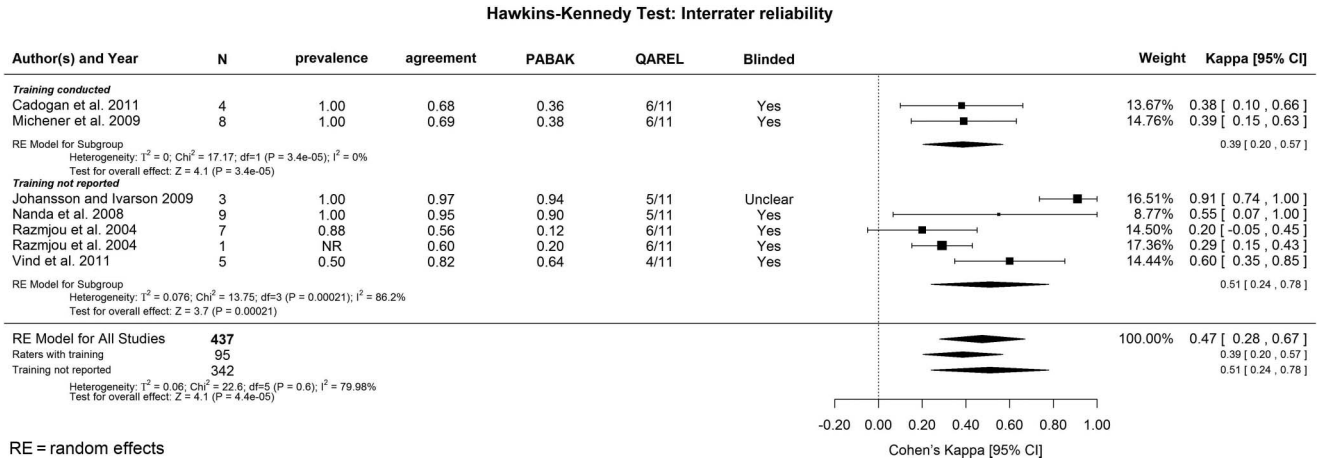

Figure 2 Hawkins-Kennedy Test. NR, not reported; PABAK, prevalence-adjusted bias-adjusted κ; QAREL, Quality Appraisal of Reliability Studies.

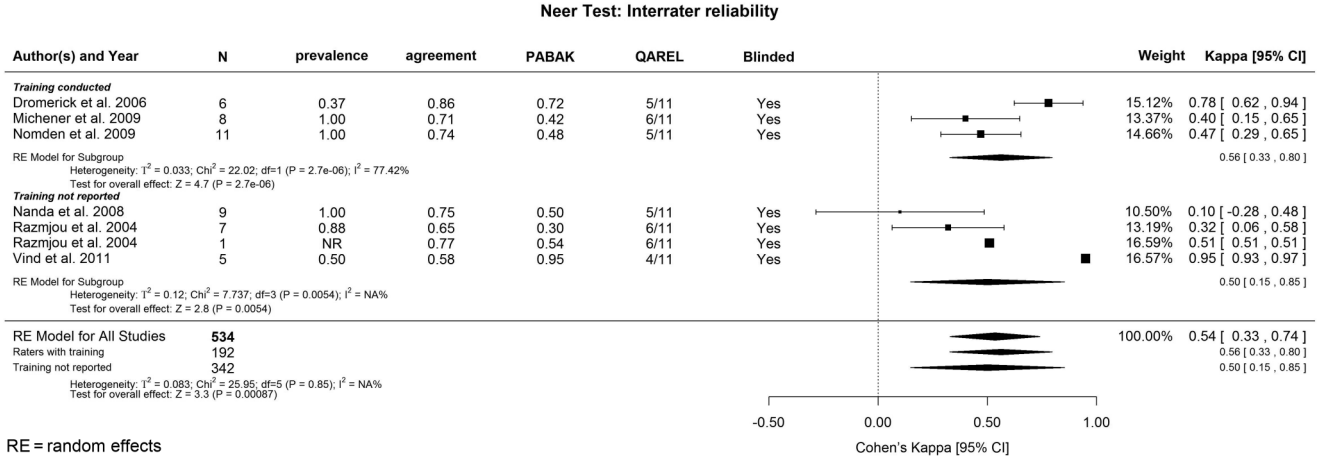

Figure 3 Neer Test. NA, not applicable; NR, not reported; PABAK, prevalence-adjusted bias-adjusted κ; QAREL, Quality Appraisal of Reliability Studies.

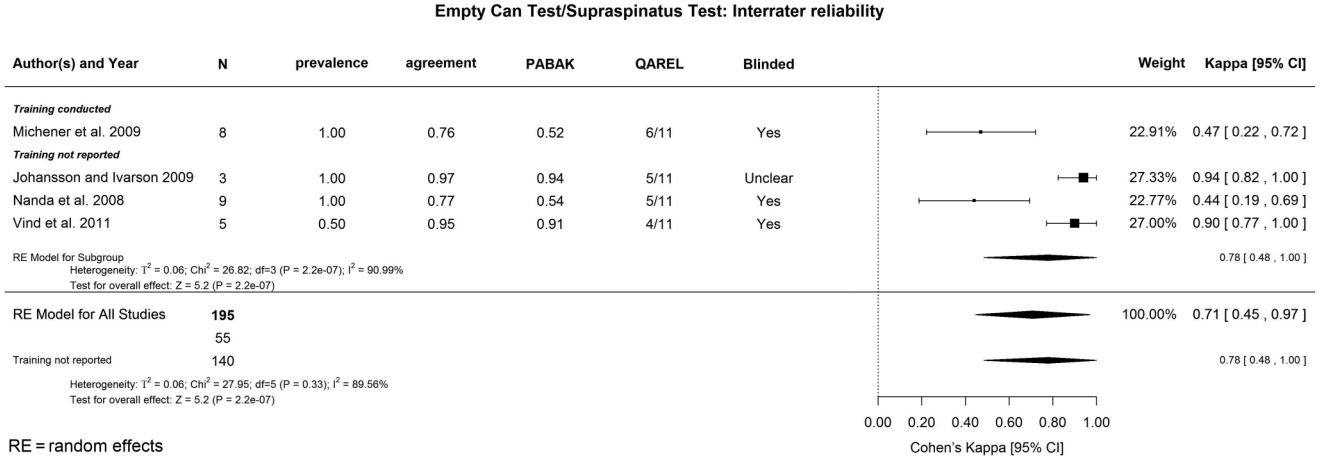

Figure 4 Empty Can Test/Supraspinatus Test. PABAK, prevalence-adjusted bias-adjusted κ; QAREL, Quality Appraisal of Reliability Studies.

diagnostic value and clinical decision-making, because of limited reliability, and also because it lacks validity.[9–11]

Physical examination tests contribute towards an overall clinical decision process that includes the patients' history, presentation and other tests and is therefore essential for clinical decision-making in patients with shoulder disorders. Physical examination manoeuvres are extensively described in the literature to be indicative of specific shoulder pathology such as rotator cuff disease, instability, and labral tears.[8–11 46–48] Prior results on diagnostic accuracy of physical examination for the shoulder are variable and therefore offer limited guidance to the clinician when assessing a patient with shoulder pain.[8 49] This

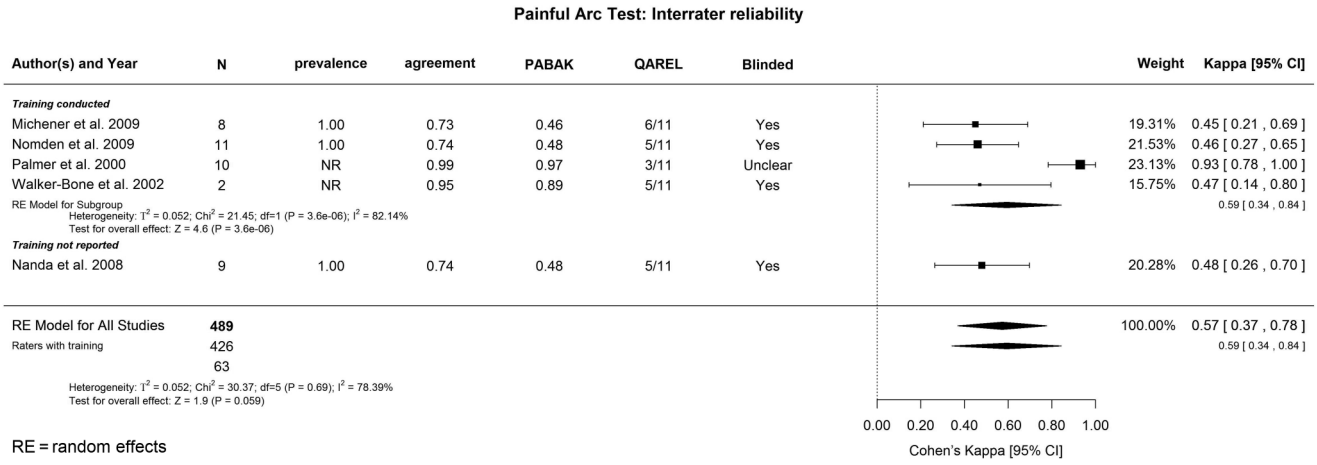

Figure 5 Painful Arc Test. NR, not reported; PABAK, prevalence-adjusted bias-adjusted κ; QAREL, Quality Appraisal of Reliability Studies.

has led to reliance on imaging for diagnostic purposes. Such practice is expensive and possibly inaccurate since imaging abnormalities are demonstrated in asymptomatic individuals as well.[50–52] Poor inter-rater and intrarater reliability is likely to be one among multiple reasons for variability in data on diagnostic accuracy when performing physical examination manoeuvres. Thus, findings from our study have implications for clinical practice and future studies on diagnostic accuracy and reliability testing. However, to perform physical examination tests should depend not only on its reliability values. It should be noted that reliable tests are not necessarily valid. For example, highly standardised tests conducted by highly trained professionals are likely more reliable in contrast to poorly standardised tests or tests conducted by untrained persons. Despite lacking standardisation or training of raters, physical examination tests may not measure the 'truth' because of multiple reasons. Thus, a high rate of false-positive (or false-negative) test results might occur, although this happens in a reliable manner. Furthermore, the validity of tests conducted by inventors of tests (or highly trained professionals) is likely not comparable to the validity of clinicians in routine care (not highly trained professionals). Hence, the reliability between this groups is inevitably lower than within the groups, because the validity depends on test performance and clinician experience.

May *et al*[18] conducted a systematic review on the reliability of physical examination tests used to assess shoulder pathologies. In contrast to this systematic review, May *et al*[18] used a self-developed tool for the quality assessment of included studies. The QAREL which was used in this systematic review, is a consensus-based developed[21] and reliable tool,[22] and has been used in recently published systematic reviews.[17 53–56]

Methodological considerations and generalisability of results

The overall generalisability of this review results is limited due to the low quality of included studies (table 1). Reliability measures reported in included studies might be inflated due to the insufficient methodology (missing blinding and randomisation of raters and patients) and statistical analysis (missing adjustment of κ values if the prevalence differs from 50%) of included studies. Highlighting this, altogether 41.63% of the QAREL items were judged as 'unclear' during critical appraisal, representing insufficient reporting of methodological aspects within primary studies. Generalisability of results from included studies is limited due to differences in test conduct as well as interpretation of physical examination tests. Since test conduct and interpretation of test results differ between studies, even if the same physical examination test was evaluated in the different studies, results from individual studies should be interpreted with caution and generalisability of such results is limited.

Rater experience and training status can have a major impact on reliability results,[13 53 57 58] but was not reported in 11 of the 18 studies included in this review.[26–30 32 35 36 40 41] Blinding of raters to the reference standard, clinical information, and additional cues was reported sufficiently in most studies.

Reliability measures may be inflated in retrospective studies, since patients might be preselected.[21] Therefore prospective studies using consecutive or randomly sampled patients should be considered for being of higher methodological quality.[17 59] However, only half of the included studies recruited patients consecutively.[25 26 29–34 37]

For the reporting of reliability study results, the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were published.[58] GRRAS intends to improve the quality of reporting, similar to the Standards for Reporting of Diagnostic Accuracy (STARD) initiative for studies of diagnostic accuracy.[60 61] To differentiate between poor reporting quality and poor methodological quality of studies is sometimes limited, but clearly poor reporting will negatively impact the proper judgement of methodological quality of studies. Therefore, some of the include studies might be judged to be of better methodological quality if reported in accordance with GRRAS. However, none of the studies included in this systematic review were published at least 1 year after the publication of GRRAS, which might be considered enough time to allow authors to use GRRAS.

An a priori sample size calculation is recommended for reliability studies,[13 62 63] but none of the studies included in this systematic review performed such a calculation (respectively no study reported that an a priori sample size calculation/post hoc power analysis was performed), potentially limiting the value and generalisability according to statistical considerations. Studies evaluating insufficient sample sizes might not be capable of producing precise estimates of agreement; therefore sample size calculations are needed and in addition will help the reader to interpret studies results.

Since prevalence rates of shoulder pathologies in routine care are presumably not equally distributed, agreement of categorical judged physical examination tests might occur purely by chance.[64] Relative reliability measures which take the agreement

occurring by chance into account, such as Cohens' κ, are therefore necessary.[62] Cohen's κ as a frequently used relative reliability measure has been criticised by several authors because Cohen's κ is affected by prevalence of test result categories.[62 65–68] If the prevalence (of the condition in the population under evaluation) differs from 50%, this will maximise the divergence between absolute (proportion of observed agreement) and relative reliability measures.[66 67] To solve this, Byrt *et al*[68] introduced the PABAK, prevalence and bias index. PABAK as a reliability measure, however, relates to a hypothetical situation without any prevalence as well as bias effects.[62] Notably, only 1[25] of the 17 studies[25–37 39–42] which reported reliability measures for categorical data provided alongside Cohen's κ values PABAK, prevalence and bias index values.

Two included studies calculated the ICC to report on the reliability of physical examination tests under evaluation;[33 38] however, in only one study[33] this was statistically appropriate since it was based on continuous data.[14 69] Measures of uncertainty and CIs were not reported in the two studies using the ICC.

It should be acknowledged that generally accepted classification systems for reliability measures are currently lacking, although the classification systems proposed Landis and Koch[20] for categorical data and Fleiss[23] for continuous data are widely used. Therefore within this review these classification systems were used to categorise the strength of agreement for individual physical examination tests. In addition, minimal requirements regarding clinical acceptable values of reliability measures are currently not available neither for categorical (eg, Cohen's κ) nor continuous data (eg, ICC),[14 16 70] but would be of great help for clinicians to decide which physical examination tests should be considered reliable for clinical use.

Implications for further research

Future reliability studies evaluating physical examination tests used for the diagnosis of shoulder pathologies should calculate and report for dichotomous outcome data contingency tables, absolute (proportion of positive as well as negative agreement) and relative reliability measures (κ, maximum κ, PABAK (all with 95% CI)), prevalence and bias index as recommended by several authors.[58 62 65–68] For continuous data, ICC values (with 95% CI) and SE of measurement should be calculated and reported.[71] The aforementioned reliability measures should be calculated and reported to enable readers to interpret, compare and adopt the reliability measures into clinical practice and research. Furthermore, reliability studies should be registered prospectively in trial registers such as the International Clinical Trials Registry Platform (ICTRP) or ClinicalTrials.gov, to ensure transparency and prospective study designs with consecutive or randomly sampled patient samples based on a priori sample size calculation should be used in reliability studies.

The reliability of physical examination test cluster(s) as described by Hegedus *et al*[4] is likely more beneficial for clinical practice in contrast to the evaluation of single tests. In addition, it seems valuable to evaluate the reliability of the use of physical examination tests only as pain or symptom-provoking procedures along with other physical movements identified by the patient that reproduce their shoulder pain as described from Lewis.[72 73]

Furthermore, an international consensus is needed regarding minimal standards for the conduct of reliability studies and reporting of studies needs to be in accordance with GRRAS.[58]

Limitations

Conclusions based on the meta-analysis results are limited due to heterogeneity and the small number of included studies. In addition, studies were included in the meta-analysis if the blinding of raters to other raters was judged as 'unclear' in the assessment of methodological quality using QAREL. This further limits interpretation of summary measures and results may be inaccurate and need to be interpreted with caution.

One study was excluded owing to language restrictions,[74] thus the possibility of a language bias might exist.

Even though authors were contacted if incomplete reliability statistics were reported in primary studies, due to several reasons not all contacted authors were able to provide the data.

CONCLUSION

Numerous physical examination tests used for the diagnosis of shoulder pathologies are described in the literature. Overall, there is a lack of high-quality studies evaluating inter-rater as well as intrarater reliability. In addition, estimates of reliability measures varied among included studies which limit conclusions that can be drawn. Despite existing heterogeneity, results from meta-analysis indicate moderate-to-substantial inter-rater reliability for the Hawkins-Kennedy Test, Neer Test, Empty Can Test/Supraspinatus Test and the Painful Arc Test. Findings from this systematic review have implications for clinical practice where physical examination manoeuvres are widely used and future studies on diagnostic accuracy and reliability testing. Evaluated physical examination tests needs to be used with great caution in terms of diagnostic value and clinical decision-making.

What are the findings?

- This is the first systematic review with meta-analysis of the reliability physical examination tests for the diagnosis of shoulder pathologies.
- Estimates of reliability measures varied among included studies which limit conclusions that can be drawn.
- Meta-analysis identified extensive heterogeneity among studies for physical examination tests, thus, the findings of the meta-analysis may be inaccurate and need to be interpreted with caution.
- Despite existing heterogeneity, results from meta-analysis indicate moderate-to-substantial inter-rater reliability for the Hawkins-Kennedy Test, Neer Test, Empty Can Test/Supraspinatus Test and the Painful Arc Test.

How might it impact on clinical practice in the future?

- Several systematic reviews have evaluated the validity of physical examination tests, concluding that most research is of insufficient methodological quality or that consistently solid measures for validity obtained from studies with higher methodological quality are lacking.
- Tests with insufficient reliability might be one reason for varying results regarding the validity of physical tests.
- The reliability of physical examination test cluster(s) is likely more beneficial for clinical practice in contrast to the evaluation of single tests.

Author affiliations
[1]Center for Evidence-Based Healthcare, University Hospital Carl Gustav Carus Dresden, Dresden, Germany
[2]Texas Tech University, School of Health Professions, Lubbock, Texas, USA
[3]International Academy of Orthopedic Medicine—Europe, Stuttgart, Germany
[4]Department of Physical Medicine and Rehabilitation, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[5]Department of Orthopaedic and Trauma Surgery, University Hospital Carl Gustav Carus Dresden, Dresden, Germany
[6]Department für Angewandte Gesundheitswissenschaften, Hochschule für Gesundheit (University of Applied Sciences), Bochum, Germany

Contributors TL made a substantial contribution to the design of the study; performed the literature search; reviewed the literature; methodologically appraised the articles; extracted, analysed and interpreted the data; produced the figures and graphs; critically revised and wrote the manuscript. OM and NBJ assisted with analysis and interpretation of data; critically revised the article and wrote the manuscript. JS and JL critically commented on the design of the study; and critically revised the manuscript. CK made a substantial contribution to the design of the study; reviewed the literature; methodologically appraised the articles; extracted the data in duplicate; analysed and interpreted the data; and critically revised and wrote the manuscript.

Funding NBJ is supported by funding from National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) 1K23AR059199.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Luime JJ, Koes BW, Hendriksen IJ, *et al*. Prevalence and incidence of shoulder pain in the general population; a systematic review. *Scand J Rheumatol* 2004;33:73–81.
2 Virta L, Joranger P, Brox JI, *et al*. Costs of shoulder pain and resource use in primary health care: a cost-of-illness study in Sweden. *BMC Musculoskelet Disord* 2012;13:17.
3 Mroz TM, Carlini AR, Archer KR, *et al*. Frequency and cost of claims by injury type from a state workers' compensation fund from 1998 through 2008. *Arch Phys Med Rehabil* 2014;95:1048–54.e6.
4 Hegedus EJ, Cook C, Lewis J, *et al*. Combining orthopedic special tests to improve diagnosis of shoulder pathology. *Phys Ther Sport* 2015;16:87–92.
5 Roy JS, Esculier JF. Psychometric evidence for clinical outcome measures assessing shoulder disorders. *Phys Thera Rev* 2013;16:331–46.
6 Calvert E, Chambers GK, Regan W, *et al*. Special physical examination tests for superior labrum anterior posterior shoulder tears are clinically limited and invalid: a diagnostic systematic review. *J Clin Epidemiol* 2009;62:558–63.
7 Munro W, Healy R. The validity and accuracy of clinical tests used to detect labral pathology of the shoulder—a systematic review. *Man Ther* 2009;14:119–30.
8 Hermans J, Luime JJ, Meuffels DE, *et al*. Does this patient with shoulder pain have rotator cuff disease?: The Rational Clinical Examination systematic review. *JAMA* 2013;310:837–47.
9 Hegedus EJ, Goode AP, Cook CE, *et al*. Which physical examination tests provide clinicians with the most value when examining the shoulder? Update of a systematic review with meta-analysis of individual tests. *Br J Sports Med* 2012;46:964–78.
10 Hegedus EJ, Goode A, Campbell S, *et al*. Physical examination tests of the shoulder: a systematic review with meta-analysis of individual tests. *Br J Sports Med* 2008;42:80–92; discussion 92.
11 Hanchard NC, Lenza M, Handoll HH, *et al*. Physical tests for shoulder impingements and local lesions of bursa, tendon or labrum that may accompany impingement. *Cochrane Database Syst Rev* 2013;(4):CD007427.
12 Sciascia AD, Spigelman T, Kibler WB, *et al*. Frequency of use of clinical shoulder examination tests by experienced shoulder surgeons. *J Athl Train* 2012;47:457–66.
13 Karanicolas PJ, Bhandari M, Kreder H, *et al*., Collaboration for Outcome Assessment in Surgical Trials (COAST) Musculoskeletal Group. Evaluating agreement: conducting a reliability study. *J Bone Joint Surg Am* 2009;91(Suppl 3):99–106.
14 Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable? *Injury* 2011;42:236–40.
15 de Vet HC, Terwee CB, Mokkink LB, *et al*. Measurement in medicine. In: de Vet HC, Terwee CB, Mokkink LB, *et al*, ed. *Measurement in medicine/a practical guide measurement in medicine: a practical guide*. Cambridge: Cambridge University Press, 2011:150–201.
16 Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;26:217–38.
17 Lange T, Freiberg A, Dröge P, *et al*. The reliability of physical examination tests for the diagnosis of anterior cruciate ligament rupture—a systematic review. *Man Ther* 2015;20:402–11.
18 May S, Chance-Larsen K, Littlewood C, *et al*. Reliability of physical examination tests used in the assessment of patients with shoulder problems: a systematic review. *Physiotherapy* 2010;96:179–90.
19 Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 2009;6:e1000100.
20 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
21 Lucas NP, Macaskill P, Irwig L, *et al*. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol* 2010;63:854–61.
22 Lucas N, Macaskill P, Irwig L, *et al*. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC Med Res Methodol* 2013;13:111.
23 Fleiss JL. *Reliability of measurement*. The Design and Analysis of Clinical Experiments: John Wiley & Sons, Inc., 1999:1–32.
24 Sun S. Meta-analysis of Cohen's kappa. *Health Serv Outcomes Res Methodol* 2011;11:145–63.
25 Cadogan A, Laslett M, Hing W, *et al*. Interexaminer reliability of orthopaedic special tests used in the assessment of shoulder pain. *Man Ther* 2011;16:131–5.
26 Dromerick AW, Kumar A, Volshteyn O, *et al*. Hemiplegic shoulder pain syndrome: interrater reliability of physical diagnosis signs. *Arch Phys Med Rehabil* 2006;87:294–5.
27 Kim SH, Ha KI, Han KY. Biceps load test: a clinical test for superior labrum anterior and posterior lesions in shoulders with recurrent anterior dislocations. *Am J Sports Med* 1999;27:300–3.
28 Kim SH, Ha KI, Ahn JH, *et al*. Biceps load test II: a clinical test for SLAP lesions of the shoulder. *Arthroscopy* 2001;17:160–4.
29 Kim SH, Park JS, Jeong WK, *et al*. The Kim test: a novel test for posteroinferior labral lesion of the shoulder—a comparison to the jerk test. *Am J Sports Med* 2005;33:1188–92.
30 Kim YS, Kim JM, Ha KY, *et al*. The passive compression test: a new clinical test for superior labral tears of the shoulder. *Am J Sports Med* 2007;35:1489–94.
31 Michener LA, Walsworth MK, Doukas WC, *et al*. Reliability and diagnostic accuracy of 5 physical examination tests and combination of tests for subacromial impingement. *Arch Phys Med Rehabil* 2009;90:1898–903.
32 Nanda R, Gupta S, Kanapathipillai P, *et al*. An assessment of the inter examiner reliability of clinical tests for subacromial impingement and rotator cuff integrity. *Eur J Orthop Surg Traumatol* 2008;18:495–500.
33 Nomden JG, Slagers AJ, Bergman GJ, *et al*. Interobserver reliability of physical examination of shoulder girdle. *Man Ther* 2009;14:152–9.
34 Norregaard J, Krogsgaard MR, Lorenzen T, *et al*. Diagnosing patients with longstanding shoulder joint pain. *Ann Rheum Dis* 2002;61:646–9.
35 Ostor AJ, Richards CA, Prevost AT, *et al*. Interrater reproducibility of clinical tests for rotator cuff lesions. *Ann Rheum Dis* 2004;63:1288–92.
36 Palmer K, Walker-Bone K, Linaker C, *et al*. The Southampton examination schedule for the diagnosis of musculoskeletal disorders of the upper limb. *Ann Rheum Dis* 2000;59:5–11.
37 Spencer DA. Mental handicap and the mental health act. *Lancet* 1977;2:502–3.
38 Tzannes A, Paxinos A, Callanan M, *et al*. An assessment of the interexaminer reliability of tests for shoulder instability. *J Shoulder Elbow Surg* 2004;13:18–23.
39 Vind M, Bogh SB, Larsen CM, *et al*. Inter-examiner reproducibility of clinical tests and criteria used to identify subacromial impingement syndrome. *BMJ Open* 2011;1:e000042.
40 Walker-Bone K, Byng P, Linaker C, *et al*. Reliability of the Southampton examination schedule for the diagnosis of upper limb disorders in the general population. *Ann Rheum Dis* 2002;61:1103–6.
41 Walsworth MK, Doukas WC, Murphy KP, *et al*. Reliability and diagnostic accuracy of history and physical examination for diagnosing glenoid labral tears. *Am J Sports Med* 2008;36:162–8.
42 Johansson K, Ivarson S. Intra- and interexaminer reliability of four manual shoulder maneuvers used to identify subacromial pain. *Man Ther* 2009;14:231–9.
43 Deeks JJ, Higgins JPT, Altman DG. Chapter 9: analysing data and undertaking meta-analyses. In: Higgins JPT, Altman DG, ed. *Cochrane handbook for systematic reviews of interventions version 5.1.0 (updated March 2011)*. The Cochrane Collaboration, http://www.cochrane-handbook.org
44 Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539–58.
45 Higgins JP, Thompson SG, Deeks JJ, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.
46 Myer CA, Hegedus EJ, Tarara DT, *et al*. A user's guide to performance of the best shoulder physical examination tests. *Br J Sports Med* 2013;47:903–7.
47 Lo IK, Nonweiler B, Woolfrey M, *et al*. An evaluation of the apprehension, relocation, and surprise tests for anterior shoulder instability. *Am J Sports Med* 2004;32:301–7.
48 Tennent TD, Beach WR, Meyers JF. A review of the special tests associated with shoulder examination. Part I: the rotator cuff tests. *Am J Sports Med* 2003;31:154–60.

49 Jain NB, Yamaguchi K. History and physical examination provide little guidance on diagnosis of rotator cuff tears. *Evid Based Med* 2014;19:108.
50 Milgrom C, Schaffler M, Gilbert S, *et al*. Rotator-cuff changes in asymptomatic adults. The effect of age, hand dominance and gender. *J Bone Joint Surg Br* 1995;77:296–8.
51 Sher JS, Uribe JW, Posada A, *et al*. Abnormal findings on magnetic resonance images of asymptomatic shoulders. *J Bone Joint Surg Am* 1995;77:10–15.
52 Yamaguchi K, Ditsios K, Middleton WD, *et al*. The demographic and morphological features of rotator cuff disease. A comparison of asymptomatic and symptomatic shoulders. *J Bone Joint Surg Am* 2006;88:1699–704.
53 Carlsson H, Rasmussen-Barr E. Clinical screening tests for assessing movement control in non-specific low-back pain. A systematic review of intra- and inter-observer reliability studies. *Man Ther* 2013;18:103–10.
54 Simopoulos TT, Manchikanti L, Singh V, *et al*. A systematic evaluation of prevalence and diagnostic accuracy of sacroiliac joint interventions. *Pain Physician* 2012;15: E305–44.
55 Rubio-Ochoa J, Benítez-Martínez J, Lluch E, *et al*. Physical examination tests for screening and diagnosis of cervicogenic headache: a systematic review. *Man Ther* 2016;21:35–40.
56 Gorgos KS, Wasylyk NT, Van Lunen BL, *et al*. Inter-clinician and intra-clinician reliability of force application during joint mobilization: a systematic review. *Man Ther* 2014;19:90–6.
57 Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Stat Med* 2002;21:3431–46.
58 Kottner J, Audigé L, Brorson S, *et al*. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011;64:96–106.
59 Fritz JM, Wainner RS. Examining diagnostic tests: an evidence-based perspective. *Phys Ther* 2001;81:1546–64.
60 Bossuyt PM, Reitsma JB. Standards for Reporting of Diagnostic Accuracy. The STARD initiative. *Lancet* 2003;361:71.
61 Bossuyt PM, Reitsma JB, Bruns DE, *et al*., Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41–4.
62 Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257–68.
63 Donner A, Rotondi MA. Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies with a binary outcome and multiple raters. *Int J Biostat* 2010;6:Article 31.
64 Gilchrist JM. Weighted 2×2 kappa coefficients: recommended indices of diagnostic accuracy for evidence-based practice. *J Clin Epidemiol* 2009;62:1045–53.
65 de Vet HC, Mokkink LB, Terwee CB, *et al*. Clinicians are right not to like Cohen's κ. *BMJ* 2013;346:f2125.
66 Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990;43:551–8.
67 Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43:543–9.
68 Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol* 1993;46: 423–9.
69 de Vet HC, Terwee CB, Knol DL, *et al*. When to use agreement versus reliability measures. *J Clin Epidemiol* 2006;59:1033–9.
70 Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? *Physiotherapy* 2000;86:94–9.
71 Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005;19:231–40.
72 Lewis J. Rotator cuff related shoulder pain: assessment, management and uncertainties. *Man Ther* 2016;23:57–68.
73 Lewis JS. Rotator cuff tendinopathy/subacromial impingement syndrome: is it time for a new method of assessment? *Br J Sports Med* 2009;43:259–64.
74 T'Jonck L, Staes F, De Smet L, *et al*. The relationship between clinical shoulder tests and the findings in arthroscopic examination. [Dutch] De relatie tussen klinische schoudertests en de bevindingen bij artroscopisch onderzoek. *Geneeskd Sport* 2001;34:15–24.

Reliability of specific physical examination tests for the diagnosis of shoulder pathologies: a systematic review and meta-analysis

Toni Lange, Omer Matthijs, Nitin B Jain, Jochen Schmitt, Jörg Lützner and Christian Kopkow

Br J Sports Med published online December 19, 2016