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

Eric J Hegedus

Correspondence to
High Point University Physical Therapy, 833 Montlieu Ave, High Point, North Carolina 27262, USA;
ehegedus@highpoint.edu

Received 21 February 2012
Accepted 21 May 2012
Published Online First 7 July 2012

ABSTRACT

Objective To update our previously published systematic review and meta-analysis by subjecting the literature on shoulder physical examination (ShPE) to careful analysis in order to determine each tests clinical utility.

Methods This review is an update of previous work, therefore the terms in the Medline and CINAHL search strategies remained the same with the exception that the search was confined to the dates November, 2006 through to February, 2012. The previous study dates were 1966 – October, 2006. Further, the original search was expanded, without date restrictions, to include two new databases: EMBASE and the Cochrane Library. The Quality Assessment of Diagnostic Accuracy Studies, version 2 (QUADAS 2) tool was used to critique the quality of each new paper. Where appropriate, data from the prior review and this review were combined to perform meta-analysis using the updated hierarchical summary receiver operating characteristic and bivariate models.

Results Since the publication of the 2008 review, 32 additional studies were identified and critiqued. For subacromial impingement, the meta-analysis revealed that the pooled sensitivity and specificity for the Neer test was 72% and 60%, respectively, for the Hawkins-Kennedy test was 79% and 59%, respectively, and for the painful arc was 53% and 76%, respectively. Also from the meta-analysis, regarding superior labral anterior to posterior (SLAP) tears, the test with the best sensitivity (52%) was the relocation test; the test with the best specificity (95%) was Yergason's test; and the test with the best positive likelihood ratio (2.81) was the compression-rotation test. Regarding new (to this series of reviews) ShPE tests, where meta-analysis was not possible because of lack of sufficient studies or heterogeneity between studies, there are some individual tests that warrant further investigation. A highly specific test (specificity >80%, LR+ ≥ 5.0) from a low bias study is the passive distraction test for a SLAP lesion. This test may rule in a SLAP lesion when positive. A sensitive test (sensitivity >80%, LR− ≤ 0.20) of note is the shoulder shrug sign, for stiffness-related disorders (osteoarthritis and adhesive capsulitis) as well as rotator cuff tendinopathy. There are six additional tests with higher sensitivities, specificities, or both but caution is urged since all of these tests have been studied only once and more than one ShPE test (ie, active compression, biceps load II) has been introduced with great diagnostic statistics only to have further research fail to replicate the results of the original authors. The belly-off and modified belly press tests for subscapularis tendinopathy, bony apprehension test for bony instability, olecranon-manubrium percussion test for bony abnormality, passive compression for a SLAP lesion, and the lateral Jobe test for rotator cuff tear give reason for optimism since they demonstrated both high sensitivities and specificities reported in low bias studies. Finally, one additional test was studied in two separate papers. The modified dynamic labral shear test, may be diagnostic of labral tears in general, but be sensitive for SLAP lesions specifically.

Conclusion Based on data from the original 2008 review and this update, the use of any single ShPE test to make a pathognomonic diagnosis cannot be unequivocally recommended. There exist some promising tests but their properties must be confirmed in more than one study. Combinations of ShPE tests provide better accuracy, but marginally so. These findings seem to provide support for stressing a comprehensive clinical examination including history and physical examination. However, there is a great need for large, prospective, well-designed studies that examine the diagnostic accuracy of the many aspects of the clinical examination and what combinations of these aspects are useful in differentially diagnosing pathologies of the shoulder.

INTRODUCTION

In 2006, we reviewed shoulder physical examination (ShPE) and in 2008 our work was published in this journal.[1] This publication was followed by a series of either similar or otherwise redundant publications, addressing all or dedicated pathognomic components of shoulder testing.[2–7] The majority of those subsequent articles did not meta-analyse the ShPE test's accuracy, evaluate risk of bias among the studies, or identify studies unique to our 2008 publication.[1] The fact that so many review articles analysed the diagnostic accuracy of clinical shoulder tests in a period of three years speaks to the need to clearly address the question. 'Which physical examination tests provide clinicians with the most value for diagnosis when examining the shoulder?'

Since 2006, there have been many changes necessitating an update of the original article. First and foremost, the publication of diagnostic articles on the use of ShPE tests in the clinical examination has continued at a brisk pace resulting in numerous new publications on the accuracy of established

tests and the development of new tests. Next, the methodology by which a systematic review on diagnostic accuracy is conducted has been updated from the Quality of Reporting of Meta-analysis (QUOROM)[8] with the publication of Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA).[9] Third, the criterion standard method of performing a meta-analysis has become a unification[10] of the bivariate model[11] and the hierarchical summary receiver operating characteristic (HSROC) model.[12] Finally, the method by which the quality of individual studies is examined has been updated from the original Quality Assessment of Diagnostic Accuracy Studies (QUADAS)[13] to the newly published QUADAS-2.[14] These changes over the last five years have been extensive but the goal with this systematic review and meta-analysis has remained the same: to analyse the literature on ShPE tests of the shoulder to careful analysis in order to determine their clinical utility in adult (18 or older) patients.

METHODS

This systematic review with meta-analysis was conducted and reported according to the protocol outlined by PRISMA[9] using a research question framed by PICOS methodology. PICOS is a pneumonic representing population (eg, adults), intervention (eg, diagnostic test), comparison (eg, control group), outcome (eg, accuracy) and study design (eg, cohort). In order to be eligible for this review, diagnostic accuracy studies, written in English, had to report both the sensitivity and specificity of ShPE tests in adults with shoulder pain due to musculoskeletal pathology. Excluded from this review, were articles using equipment or devices that are not readily available to most clinicians during physical examination and articles in which subjects were tested under anaesthesia or in which subjects were cadavers.

Study selection

Since this review is an update of our previous work,[1] the terms in our Medline and CINAHL search strategies remained the same with the exception that the search was confined to the dates November, 2006 through February, 2012. Our previous study dates were 1966 – October, 2006. Further, the original search was expanded, without date restrictions, to include two new databases: EMBASE and the Cochrane Library. A hand search was also conducted which included the authors' private collections and the searching of previous systematic reviews. Two authors (EH and AW) read titles and abstracts of all database-captured articles applying the a priori inclusion/exclusion criteria and agreement was measured using the κ statistic (figure 1). Disagreement was then resolved by discussion between the two authors and, in the event that agreement could not be reached, a third author (CC) served as the deciding vote. With the remaining articles, the same two authors (EH and AW) read the entire paper and again, a κ value was calculated to measure agreement as to which articles to retain for final analysis (figure 1). Once the final group of 32 articles was determined, 2x2 table data were extracted and saved for meta-analysis. Only data from studies, where the 2x2 data were reported or could be inferred from stated positive likelihood ratios, negative likelihood ratios, positive predictive values, and negative predictive values were retained for meta-analysis. If 2x2 data could not be discerned, the article was excluded from meta-analysis but still retained for systematic review and qualitative analysis.

Quality assessment

Once the final group of articles was agreed upon, two authors (EH and AW) independently examined the quality of each article using the QUADAS-2 tool.[14] QUADAS-2 is a 4-phase tool, the last phase of which assists authors of systematic reviews in rating: 1) bias and 2) applicability. The risk of bias is assessed in four key areas: patient selection, index test, reference standard, and flow and timing. Concern for applicability is assessed in three key areas: patient selection, index test, and reference standard. For both categories, risk of bias and concern for applicability, the individual criteria were classified as low risk, high risk, or unclear and the results were presented using tables from the QUADAS web site (www.quadas.org).

Statistical analysis

In order to maximise the potential for meta-analysis, we added 2x2 data from our first meta-analysis[1] to data gathered from the 32 additional articles included in this review. Hierarchical summary receiver operating characteristic (HSROC) curve[12] and bivariate[11] models were used to combine estimates of sensitivity (SN), specificity (SP), positive likelihood ratios (+LR), negative likelihood ratios (–LR) and diagnostic OR (DOR) with their 95% CI. Sensitivity measures the proportion of actual positives which are correctly identified as such (eg, the percentage of sick people who are correctly identified as having the condition). Specificity measures the proportion of negatives which are correctly identified (eg, the percentage of healthy people who are correctly identified as not having the condition). Positive likelihood ratio (LR+) dictates how much the odds of the disease increase when a test is positive.[15] The negative likelihood ratio (LR–) dictates how much the odds of the disease decrease when a test is negative.[15] Diagnostic OR express the strength of association between the test result and disease. These models, in the absence of covariates, are different parameterisations of the same model[10] and take into account the correlation between sensitivity and specificity and both the within and the between study variances.[16] The 95% prediction region is graphically provided which is the given probability (ie, 95%) of including the true sensitivity and specificity of a future study.[17] DerSimonian-Laird[18] random-effects models were used where less than four studies were eligible for statistical pooling. Heterogeneity was explored graphically with forest plots and statistically with Cochrane-Q with $p<0.10$ to indicate significant heterogeneity. When appropriate, meta-regression or subgroup analysis using study level characteristics was used to explore heterogeneity with a $p<0.10$ to indicate a significant difference in stratified estimates. A p value of <0.10 was decided upon to determine a significance in stratified estimates due to the low power of the test used to detect differences in stratified estimates.[19] A 0.5 was added to all four cells of the 2x2 table when a zero was encountered in any cell as suggested by Cox.[20]

Publication bias was analysed statistically with the Egger[21] test with a $p<0.05$ to indicate significant publication bias. Threshold effects were tested using Spearman correlation coefficients.[22] Influential studies on summary estimates were assessed with Cooks-d and standardised residuals according to Rabe-Hesketh[23] with sensitivity analyses to determine if influential studies should be removed from the analyses. All statistical analyses were conducted in Stata 11 (Stata, College Station Texas, USA) by one of the authors (AG).

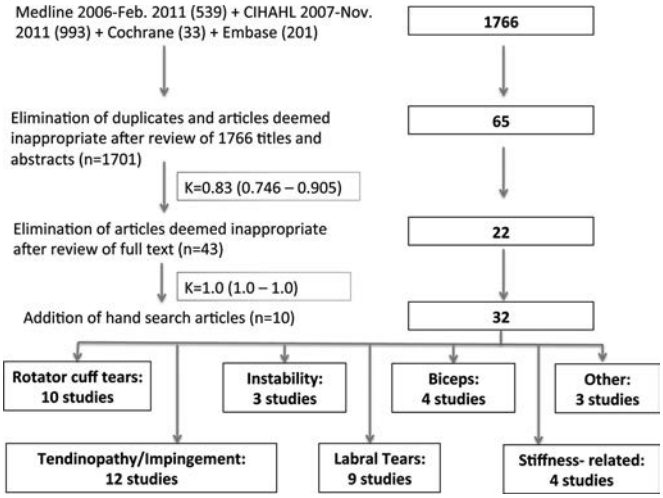

Figure 1 Flow diagram of the literature screening process. Note that the total of articles broken down into subgroups does not equal 32 because multiple articles addressed more than one pathognomonic category. This figure is only reproduced in colour in the online version.

RESULTS

New Studies/Tests/Pathologies

In reference to our previous meta-analysis,[1] there were 32 new studies addressing the diagnostic accuracy of ShPE tests of the shoulder in adults (figure 1). A summary of the characteristics of each study is presented in table 1.

Twelve of these studies[26 28 29 35 38 39 45–49 53] added 13 new tests to the literature, the majority of which attempted to detect a SLAP lesion. New tests were defined as those for which diagnostic accuracy statistics were reported for the first time in peer-reviewed literature. Clinically, many of these tests are not new. The 32 studies addressed the categories of: Rotator cuff tears (RCT's), Tendinopathy, Subacromial impingement, Instability, Labral tears, Biceps pathology, Stiffness-related disorders and Other. The most frequent topics of focus were RCTs, Tendinopathy, Subacromial impingement and Labral tears. Many would consider tendinopathy and impingement different labels for the same syndrome and further, that both labels capture a continuum of disease that includes RCTs. We concur with this thought but separated these pathologic entities in order to simplify analysis. Therefore, the rotator cuff tear group included those studies where diagnostic accuracy was examined inclusive of any size of tear or classification system used. Three studies[25 30 33] in the RCT category addressed full-thickness tears, one study[39] addressed massive RCTs, and six studies[41 42 46 52–54] addressed RCT's regardless of size or classification. Of the 10 RCT studies, five used tests designed to test specific, individual muscles of the rotator cuff. An example of this methodology was the Kim *et al*[42] study that examined the accuracy of the empty can for supraspinatus pathology, Patte's test for infraspinatus tendinopathy, and the lift-off for subscapularis tendinopathy (and Yergason's test for biceps tendinopathy).

There were some trends observed in categories other than RCTs. In the labral tear group, two studies examined the use of tests to detect any labral tear, while six studies addressed superior labral anterior to posterior (SLAP) lesions and one study[37] addressed both labral tears generally and SLAP lesions specifically. Of the three studies in the Instability category,[29 37 39] one[39] addressed soft tissue-related instability and two[29 37] addressed bony instability, a pathology attracting increased attention since our last review. The Stiffness-related group included studies addressing either glenohumeral OA or adhesive capsulitis. Two studies[28 39] in this category actually used the same data for the shrug sign and published that data in two separate papers. All three of the stiffness-related papers[28 39 48] addressed adhesive capsulitis, another new pathology in the diagnostic literature since our last review. Finally, the Other category consists of two articles[38 39] on detecting acromioclavicular (AC) pathology and one addressing bony abnormality.[47]

The sensitivity and specificity of most ShPE tests examined in all 32 studies and the risk of bias in each study are summarised in table 2. In the interest of efficient reporting, test data was omitted from table 2 if diagnostic accuracy figures were reported for pathologies which the test was never intended to detect. For example, if an author reported values for the lift-off test (subscapularis) in a population with adhesive capsulitis, that data were not reported.

Quality assessment – risk of bias and concern for applicability

Each of the 32 papers qualifying for final review was scrutinised, via the QUADAS-2 (Q2),[14] in the areas of risk of bias and concern for applicability (Appendix). Concern for applicability, for this review, was defined as concern for external validity, the degree to which results of a research study can be applied to practice. The two authors (EH and AW) independently used the Q2,[14] blinded from each other's assessments. The number of low risk/concern scores was tallied into a total score for each article and agreement was calculated using a weighted κ statistic. The weighted κ was poor (κ=0.31 with 95% CI 0.10 to 0.52). Summaries of risk of bias and concern for applicability for each pathological group are presented in figure

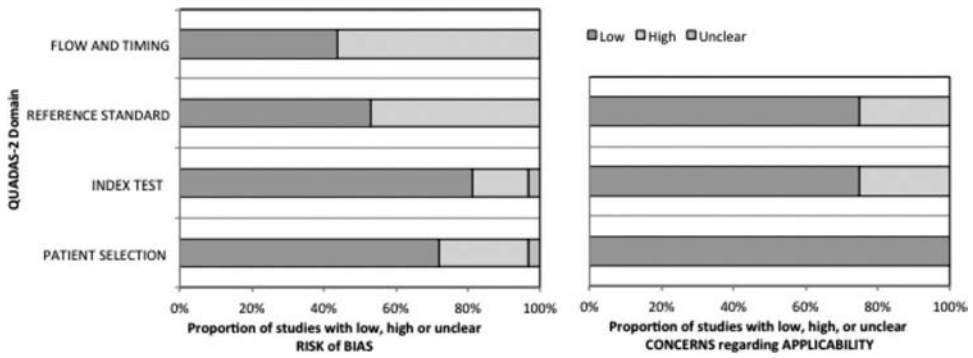

Figure 2 Risk of bias and concerns for applicability. Green=low risk/concern; Orange=high risk/concern; Blue=uncertain risk/concern. This figure is only reproduced in colour in the online version.

Table 1 Summary of studies

Lead author, year	Mean age years (range)	Mean symptom duration	Study design	Criterion standard	ShPE test	LR±	LR−	Author conclusions
Michener[24]	40.6 (18–83)	33.8 months	Prospective blinded study	Arthroscopy	Hawkins-Kennedy	1.63	0.61	The single tests of painful arc, external rotation resistance, and Neer are useful screening tests to rule out SAIS (subacromial impingement syndrome). The reliability of all tests was acceptable for clinical use. Based on reliability and diagnostic accuracy, the single tests of the painful arc, external rotation resistance, and empty can have the best overall clinical utility. The cut point of 3 or more positive of 5 tests can confirm the diagnosis of SAIS, while less than 3 positive of 5 rules out SAIS.
					Neer	1.76	0.35	
					Painful arc	2.25	0.38	
					Empty can	3.90	0.57	
					Resisted External rotation/ Infraspinatus test	4.39	0.50	
					3 or more positive tests	2.93	0.34	
Miller[25]	55.5 (20–86)	37.5 months	Case-control, same subject, correlation, double-blind	Ultrasound	External rotation lag sign	7.2	0.60	A positive sign would appear to suggest the moderate likelihood of the presence of a full-thickness tear but this conclusion is tenuous based on the small sample size of the study and subsequent wide confidence interval
					Drop sign	3.2	0.30	
					Internal rotation lag sign	6.2	0.00	
Kim YS[26]	32.6 (19–54)	NA	Cohort study	Arthroscopy	Passive compression test	5.9	0.21	The passive compression test is a useful and accurate technique for predicting superior labral tears of the shoulder joint.
Fodor[27]	57 (20–84)	NA	Prospective, consecutive subjects	Ultrasound	Neer	10.8	0.48	The Hawkins-Kennedy test is the most sensitive test for identification of subacromial impingement syndrome, while Neer is most specific. With 4 (+) tests, the specificity increases and the sensitivity decreases. No tests were good at distinguishing stages of subacromial impingement.
					Yocum	8.80	0.33	
					Hawkins-Kennedy	6.50	0.31	
					Painful arc	3.40	0.41	
					4 tests	1.70	0.03	
Jia[28]	NA	NA	Retrospective	Arthroscopy	Shrug sign			The shrug sign is a non-specific physical exam sign for shoulder dysfunction and is more commonly associated with glenohumeral OA, adhesive capsulitis, and massive rotator cuff tear.
					Glenohumeral OA	3.6	0.12	
					Adhesive capsulitis	1.90	0.10	
					Massive RC tear	1.50	0.50	
					Rotator cuff tendinopathy	2.04	0.08	
					FTT RC	1.30	0.72	
Bushnell[29]	24 (16–52)	NA	prospective pilot study	Arthroscopy	Bony apprehension test for instability	7.14	0.00	The bony apprehension test can reliably screen for significant osseous lesions.
Castoldi[30]	50.4 (16–89)	NA	Prospective cohort treatment	Arthroscopy	External rotation lag sign (ERLS) – FTT SS	28.00	0.45	The ERLS is highly specific and acceptably sensitive for the diagnosis of full-thickness tears, even in case of an isolated lesion of the supraspinatus tendon.
						13.86	0.03	
					ERLS – FTT SS & IS	14.29	0.00	
					ERLS – FTT TM			
Silva[31]	55 (24–82)	97.5 days	Prospective	MRI	Neer	0.98	1.10	The Yocum test was the most sensitive for subacromial impingement and the Gerber test for subacromial-subdeltoid bursitis. The Gerber and Patte tests provide the best diagnostic combo. The majority of tests showed low specificity.
					Hawkins-Kennedy	1.23	0.65	
					Yocum	1.32	0.53	
					Jobe	1.06	0.87	
					Patte	1.50	0.67	
					Gerber	1.36	0.64	
					Resisted abduction	0.73	2.10	
Chew[32]	44 (18–75)	9.8 months	prospective cohort	Ultrasound	Neer	1.60	0.60	Diagnosis of supraspinatus pathology may be accomplished with a cluster of three tests: age >39, (+) painful arc, self reported clicking or popping.
					Hawkins-Kennedy	1.30	0.40	
					Cross body adduction	1.90	0.40	
					Drop arm	3.30	0.80	
					Full can	2.40	0.40	
					Empty can	1.60	0.30	
					Painful arc	3.70	0.40	
Bak[33]	56 FTT, 38 No tear (39–75 FTT; 19–73 control)	13 days	Prospective diagnostic study	Ultrasound	Hawkins-Kennedy	1.04	0.88	BEFORE subacromial lidocaine injection: external rotation lag sign or drop arm test are indicative of a FTT supraspinatus; negative lag sign does not preclude a tear. AFTER subacromial lidocaine injection: specificity improves and sensitivity is reduced for all tests.
					Neer	0.92	1.14	
					Jobe	1.25	0.62	
					Painful arc	100	1.00	
					Drop arm test	2.41	0.71	
					External rotation lag sign	5.00	0.60	
					Infraspinatus drop sign	1.50	0.79	
					Internal rotation lag sign	2.38	0.79	

Table 1 *Continued*

Lead author, year	Mean age years (range)	Mean symptom duration	Study design	Criterion standard	ShPE test	LR±	LR−	Author conclusions
Bartsch[34]	58 (SD 11.6)	NA	Prospective, consecutive subjects	Arthroscopy	Lift off test	1.90	0.76	Fifteen percent of the subscapularis tears were not predicted preoperatively by using all of the tests. The modified BPT and the BOS showed the greatest sensitivity. The BOS had the greatest specificity. With the BOS and the modified BPT in particular, upper subscapularis lesions could be diagnosed preoperatively.
					Internal rotation lag sign	1.30	0.64	
					Modified belly press test (BPT)	2.75	0.18	
					Belly off sign (BOS)	9.67	0.14	
Kibler[35]	49 (28–64)	NA	Cohort study	Arthroscopy		biceps/SLAP	biceps/SLAP	The upper cut test shows higher levels of clinical utility for the detection of biceps injuries than traditional tests. The likelihood ratio, however, suggest its individual value is moderate. Therefore, the upper cut & Speed's tests together provide fairly high clinical prediction of arthroscopic biceps pathology. The modified dynamic labral shear test shows the highest level of clinical utility in the diagnosis of SLAP tears when compared to any individual tests. The modified dynamic labral shear test & O'Brien's together show best prediction of SLAP arthroscopic findings.
					Belly press	2.1/.61	0.81/1.13	
					Upper cut	3.38/.49	0.34/1.40	
					Bear hug	1.95/.54	0.36/1.98	
					Yergason's	1.94/.88	0.74/1.05	
					Speed's	2.77/.93	0.58/1.03	
					Dynamic labral shear test	0.38/31.57	1.54/.29	
					Anterior slide	0.64/2.63	1.22/0.64	
					O'Brien's	0.96/3.83	1.02/.47	
Chen[36]	NA	23 weeks	Prospective, double-blind	Ultrasound	Yergason's	1.47	0.87	All three tests are limited by poor sensitivity with respect to biceps tendinosis.
					Speed's	1.55	0.63	
					Bicipital groove palpation	2.04	0.60	
Fowler[37]	41	35 weeks	Cohort/retrospective	Arthroscopy	Hawkins-Kennedy/RC tendinopathy	2.10	0.60	The diagnostic accuracy of isolated standard shoulder tests in recreational athletes was overall very poor. A positive response gained in one of a combination of clinical tests caused test sensitivity to increase substantially in all pathological conditions, with specificity subsequently plummeting.
					Relocation/Bankart lesion	6.10	0.20	
					Relocation/Hill-Sachs	4.30	0.20	
					O'Brien's/labrum	1.05	0.90	
					O'Brien's/SLAP	1.10	0.80	
					Apprehension/SLAP	1	1	
					Gerber's/RC tendinopathy	1.9	0.9	
Goyal[38]	45 (23–62)	2.8 months	Case-control	Ultrasound	Speed's	4.6	0.34	Sensitivity was good in the clinical diagnosis of supraspinatus lesions and low in other shoulder lesions, especially the infraspinatus and the acromioclavicular joint. Specificity was high for lesions of infraspinatus, subscapularis and the acromioclavicular joint. However, it was fairly good for biceps tendon pathology and very low for the supraspinatus lesions. Physical exam was unable to differentiate rotator cuff tendonitis from tear, and partial-thickness tears from full-thickness tear.
					Resisted abduction	1.43	0.26	
					Resisted external rotation	NA	0.5	
					Resisted internal rotation	26.78	0.26	
					Adduction stress	15	0.45	
Jia[39]	NA	NA	Retrospective	Arthroscopy	AC resisted extension	4.8	0.33	The results of shoulder examinations are variable and statistical analysis may not demonstrate a substantial improvement on the original observations of Codman.
					Active compression for SLAP	1.26	0.81	
					Active compression for biceps tendinopathy	1.26	0.70	
					Active compression for AC joint OA	8.20	0.62	
					Anterior Slide	2.63	0.62	
					Anterior apprehension for glenohumeral instability	14.50	0.44	
					Anterior apprehension for anterior instability	18.00	0.29	
					Posterior apprehension	19.00	0.82	
					Cross body for RC tendinopathy	0.88	1.04	
					Cross body for AC joint OA	3.67	0.29	
					Drop arm	2.18	0.39	
					External rotation lag sign for massive RC tear	3.18	0.73	
					External rotation lag sign for RC tendinopathy	0.44	1.11	

Table 1 *Continued*

Lead author, year	Mean age years (range)	Mean symptom duration	Study design	Criterion standard	ShPE test	LR±	LR−	Author conclusions
					Hawkins-Kennedy for AC joint OA	0.85	1.18	
					Hawkins Kennedy for biceps tendinopathy	0.89	1.18	
					Lift off for biceps tendinopathy	2.55	0.81	
					Lift off for glenohumeral OA	2.90	0.79	
					Lift off for RC tendinopathy	0.48	1.14	
					Neer for AC joint OA	0.97	1.05	
					Neer for biceps tendinopathy	1.08	0.88	
					Speed's for biceps tendinopathy	1.52	0.75	
					Whipple for massive RC tear	1.35	0.00	
					Whipple for RC tendinopathy	1.19	0.61	
Kelly[40]	57 (20–70)	2 years	Cross-sectional study	Ultrasound	Neer	0.62		The Hawkins-Kennedy test was the most accurate test for diagnosing any degree of subacromial impingement syndrome. The most accurate tests for diagnosing sub-categories of impingement were pain on resisted external rotation and weakness during the full can test for presence of subdeltoid fluid, pain on resisted external rotation for partial-thickness tears and the painful arc test for full-thickness tears. Overall, physical tests have limited diagnostic value.
					Hawkins-Kennedy	1.48	0.52	
					Painful arc of abduction	0.59	1.4	
					Abduction weakness	0.76	1.24	
					Abduction pain	2.21	0.34	
					External rotation weakness	0.74	1.8	
					External rotation pain	3.3	0.74	
					Empty can weakness	1.56	0.72	
					Empty can pain	0.78	1.45	
					Full can weakness	1.79	0.73	
					Full can pain	0.46	2.6	
Kim HA[41]	59 (37–77)	16.1 months	Prospective	Ultrasound	Empty can (SS)	0.64	1.34	Physical examination used for the diagnosis of shoulder pain had low sensitivity and specificity for the detection of rotator cuff tendon tears.
					Lift -off (SB)	0.081	4.67	
					Yergason's (biceps)	4.03	0.31	
Kim HA[42]	53 (16–75)	10 months	Prospective	Ultrasound	Empty can (SS)	1.3	0.6	Physical examination of the rotator cuff and biceps had low sensitivity and specificity in the rheumatoid shoulder joint.
					Patte's (IS)	2.3	0.5	
					Lift -off (SB)	1.3	0.7	
					Yergason's (biceps)	1.3	0.96	
Salaffi[43]	58 (23–81)	2 months	Prospective, consecutive subjects	Ultrasound	Hawkins-Kennedy	2.15	0.51	The sensitivity was low for the clinical diagnosis of all shoulder abnormalities. As calculated through an ROC curve analysis, the Simple Numeric Pain by SHOulder Test (SNAPSHOT) index may improve the clinical examination of the painful shoulder by overcoming the low clinical value of each single maneuver. The SNAPSHOT optimal cut-off point was a score of >3 which increased the specificity and likelihood ratios considerably.
					Empty can	1.14	0.85	
					Patte's test	2.43	0.5	
					Lift –off	1.45	0.85	
					Speed's	2.1	0.66	
					SNAPSHOT >3	6.1	0.3	
Walsworth[44]	40 (18–83)	34	Prospective cohort	Arthroscopy	Active compression	0.67	2.5	The combination of popping or catching with a positive crank or anterior slide result or a positive anterior slide result with a positive active compression or crank test result suggests the presence of a labral tear. The combined absence of popping or catching and a negative anterior slide or crank result suggests the absence of a labral tear.
					Anterior slide	2.38	0.69	
					Crank	1.35	0.71	
Schlecter[45]	44 (13–84)	NA	Retrospective analysis	Arthroscopy	Passive distraction test (PDT)	8.83	0.5	The passive distraction test can be used alone or in combo to aid the clinical evaluation and diagnosis of SLAP lesion.
					Active compression test (ACT)	7.38	0.45	
					Anterior slide test	10.5	0.81	
					PDT + ACT	7	0.33	

Table 1 *Continued*

Lead author, year	Mean age years (range)	Mean symptom duration	Study design	Criterion standard	ShPE test	LR±	LR−	Author conclusions
Gillooly[46]	53 (17–83)	15	Prospective cohort	Arthroscopy	Lateral Jobe test	7.36	0.21	The lateral Jobe test had a higher sensitivity than the combined tests (Empty can, strength in ER, and subacromial impingement tests). It is a simple, new technique which can improve the clinical diagnosis of rotator cuff tears; *positive result for the combined tests was taken as weakness on supraspinatus testing, weakness in external rotation and pain on subacromial impingement or a combination of two of these and an age greater than 60 years.
					Combined tests*	4.75	0.49	
Adams[47]	NA	acute	Prospective	X-rays	Olecranon-Manubrium Percussion Sign		0.15	The presence of a normal OMP (olecranon manubrium percussion test) sign does not negate the need for radiographic studies in patients with shoulder injury. The presence of an abnormal OMP sign suggests the need for appropriate radiographic studies.
Carbone[48]	40–50	NA	Retrospective	Codman's criteria, exam &/or MRI	Coracoid pain test (Adhesive capsulitis)	49.5	0.01	Coracoid pain test is an easy and reliable test for identifying patients with or without adhesive capsulitis.
Ebinger[49]	49 (14–79)	chronic	Prospective	Arthroscopy	Supine flexion resistance test	2.6	0.29	Regarding type II SLAP lesions, the supine flexion resistance test is more specific than the O'Brien's or Speed's test.
					Speed's	0.97	1.05	
					Active compression	1.3	0.21	
Cook[50]	45	chronic	Prospective, case-based, case-control	Arthroscopy	Active compression	1.1	0.67	Each of the 5 stand-alone tests and clusters of tests provide minimal to no value in the diagnosis of a SLAP lesion, whether a SLAP-only lesion or a SLAP lesion with or without a concomitant findings reference.
					Kim II	1.2	0.85	
					Dynamic labral shear	1.3	0.4	
					Speed's	1.1	0.94	
					Labral tension	1.2	0.94	
Gill[51]	44 no tear/59 partial tear	NA	Cohort study	Arthroscopy	Palpation	1.13	0.87	No single physical examination test can accurately predict the presence of a partial tear of the long head of the biceps tendon.
					Lift off	2.61	0.81	
					Speed's	1.51	0.75	
Kim, E[52]	60 (37–83)	>3 months	Prospective	MRI and arthroscopy	Empty can pain or weakness for RC tear (PTT or FTT)	1.74	0.02	Both the empty can test and full can test were considered to be valuable as screening tests to detect a torn rotator cuff, using the positive signs of pain and weakness separately, in spite of their modest overall accuracy.
					Empty can weakness for RC tear (PTT or FTT)	2.62	0.34	
					Empty can pain for RC tear (PTT or FTT)	1.74	0.13	
					Empty can pain and weakness for RC tear (PTT or FTT)	2.73	0.39	
					Full can pain or weakness for RC tear (PTT or FTT)	1.96	0.19	
					Full can weakness for RC tear (PTT or FTT)	2.41	0.34	
					Full can pain for RC tear (PTT or FTT)	2.22	0.43	
					Full can pain and weakness for RC tear (PTT or FTT)	3.28	0.50	
Naredo[53]	58 (21–77)	12.5 months	Prospective	Ultrasound	Empty can pain or weakness, SS tear	Infinity	0.81	The accuracy of clinical diagnosis of periarticular shoulder conditions is low. Physical exam was unable to differentiate rotator cuff tendinitis from tear, and partial thickness tear from full-thickness tear.
						1.58	0.42	
					Empty can pain or weakness, SS tendinopathy	3.1	0.6	
						7.1	0.3	
					Lift-off, SB tendinopathy	7.2	0.67	
					Patte's, IS tendinopathy			
					Patte's, IS tear			

Table 1 *Continued*

Lead author, year	Mean age years (range)	Mean symptom duration	Study design	Criterion standard	ShPE test	LR±	LR−	Author conclusions
Itoi[54]	53 (16–86)	NA	Retrospective case series	Arthroscopy	SS tear: Full can pain	1.6	0.4	Pain is not useful in locating the sight of a tear. In patients with cuff tendinopathy, the supraspinatus test is most accurate when interpreted with MMT < 5, whereas ERST (infraspinatus) is most accurate with MMT < 4+, and lift-off test (subscapularis) most accurate with MMT <3
					SS tear: Full can weak (MMT<5)	1.8	0.32	
					SS tear: Empty can pain	1.3	0.55	
					SS tear: Empty can weak (MMT<5)	1.5	0.3	
					IS tear: External rotation strength test pain	1.17	0.9	
					IS tear: External rotation strength test weak < 5	1.8	0.3	
					SB tear: Lift-off test pain	1.5	0.8	
					SB tear: Lift-off test weak <5	1.9	0.4	
Oh[55]	Mid 40's (17–mid 70's)	NA	Retrospective case control study	Arthroscopy	Biceps groove tenderness	1.7	1.1	No test had a high sensitivity and high specificity; no combination of 2 tests yielded sensitivity/specificity of more than 60%. Combinations of 2 sensitive tests (O'Brien's, Anterior apprehension, Compression rotation) and 1 specific test (Speed's, Yergason, biceps load II) increased the diagnostic accuracy. Requiring 1 of 3 tests to be positive, will result in a sensitivity of ~ 75%, whereas all 3 positive results in a specificity of ~ 90%.
					Speed's	1.06	1	
					Yergason	0.92	1	
					Relocation	0.96	1	
					Compression-rotation	1.3	0.72	
					Active compression	1.3	0.7	
					Kibler	0.7	1.1	
					Biceps load II	1.4	0.9	
					Anterior apprehension	1.1	0.9	
					Whipple	1.1	0.8	

AC, acromioclavicular; FTT, full-thickness tear; IS, infraspinatus; MMT, manual muscle test; NA, not available; OA, osteoarthritis; PTT, partial-thickness tear; RC, rotator cuff; SB, subscapularis; SLAP, superior labrum anterior posterior; SS, supraspinatus; TM, teres minor.

Table 2 Alphabetical list of common shoulder physical examination (ShPE) tests

Test name(s)	Pathology	Lead Author	Sensitivity	Specificity	Risk of Bias* from QUADAS 2
AC Resisted Extension	AC Joint OA	Jia[39]	72	85	High
Active Compression/O'Brien	SLAP	Cook[50]	91	14	Moderate
	SLAP	Schlecter[45]	59	92	Low
	SLAP	Ebinger[49]	94	28	Low
	Type II SLAP	Oh[55]	63	53	Low
	SLAP	Jia[39]	53	58	High
	Labral Tear	Kibler[35]	61	84	Low
	Labral Tear	Fowler[37]	63	40	High
	SLAP	Fowler[37]	64	43	High
	Biceps Tendinopathy	Kibler[35]	38	61	Low
	Biceps Tendinopathy	Jia[39]	68	46	High
	AC Joint OA	Jia[39]	41	95	High
	Labral Tear	Walsworth[44]	55	18	Low
Adduction Stress	AC joint OA	Goyal[38]	57	96	High
Anterior Slide	Biceps Tendinopathy	Kibler[35]	24	62	Low
	Biceps Tendinopathy	Jia[39]	50	81	High
	Labral Tear	Kibler[35]	48	82	Low
	SLAP	Schlecter[45]	21	98	Low
	Type II SLAP	Oh[55]	21	70	Low
	Labral Tear	Walsworth[44]	43	82	Low
Apprehension- Anterior	Type II SLAP	Oh[55]	62	42	Low
	SLAP	Fowler[37]	29	70	High
	SLAP	Fowler[37]	29	70	High
	Glenohumeral Instability	Jia[39]	58	96	High
	Anterior Instability	Jia[39]	72	96	High
Apprehension- Posterior	Posterior Instability	Jia[39]	19	99	High
Bear Hug	Biceps Tendinopathy	Kibler[35]	79	60	Low
	Labral Tear	Kibler[35]	37	32	Low
Belly-off	Subscapularis Tendinopathy	Bartsch[34]	86	91	Low
Belly Press	Biceps Tendinopathy	Kibler[35]	31	85	Low
	Labral Tear	Kibler[35]	15	75	Low
Belly Press (modified)	Subscapularis Tendinopathy	Bartsch[34]	80	88	Low
Belly Press (resisted)	Subscapularis Tendinopathy	Goyal[38]	75	97	High
Biceps Load II	SLAP	Cook[50]	55	53	Moderate
	Type II SLAP	Oh[55]	30	78	Low
Bony Apprehension	Bony Instability	Bushnell[29]	94	84	Low
Compression-Rotation	Type II SLAP	Oh[55]	61	54	Low
Crank	Labral Tear	Walsworth[44]	61	55	Low
Cross-body	Supraspinatus Tendinopathy	Chew[32]	75	61	Low
	RC Tendinopathy	Jia[39]	22	75	High
	AC Joint OA	Jia[39]	77	79	High
Drop-arm	Supraspinatus Tendinopathy	Chew[32]	24	93	Low
	FTT- Supraspinatus	Bak[33]	41	83	High
	RC Tendinopathy	Jia[39]	74	66	High
Drop Sign	FTT- Supraspinatus	Bak[33]	45	70	High
	FTT- Supraspinatus/Infraspinatus	Miller[25]	73	77	Moderate
Dynamic Labral Shear	SLAP	Cook[50]	89	30	Moderate
Dynamic Labral Shear- Modified	Labral Tear	Kibler[35]	72	98	Low
	Biceps Tendinopathy	Kibler[35]	18	53	Low
Empty Can (pain)	Torn Supraspinatus	Itoi[54]	78	40	Moderate
	Subacromial impingement	Kelly[40]	52	33	Low
	RC Tear	Kim E[52]	94	46	Moderate
Empty Can (weak)	Torn Supraspinatus	Itoi[54]	87	43	Moderate
	Subacromial impingement	Kelly[40]	52	67	Low
	RC Tear	Kim E[52]	76	71	Moderate
	Subacromial impingement	Michener[24]	50	87	Low
Empty Can (pain or weak)	Supraspinatus Tendinopathy	Chew[32]	83	49	Low
	RC Tear	Kim E[52]	99	43	Moderate
	FTT- Supraspinatus	Bak[33]	76	39	High
	Supraspinatus Tear	Naredo[53]	19	100	Moderate
	Supraspinatus Tendinopathy	Kim HA[42]	72	45	High
	Supraspinatus Tendinopathy	Kim HA[41]	31	52	Low
	Supraspinatus Tendinopathy	Fodor[27]	50	83	Moderate
	Supraspinatus Tendinopathy	Salaffi[43]	56	51	Moderate
	Supraspinatus Tendinopathy	Naredo[53]	79	50	Moderate
	Supraspinatus Tendinopathy	Goyal[48]	90	37	High
Empty Can (pain and weak)	Subacromial impingement	Silva[31]	74	30	Low
	RC Tear	Kim E[52]	71	74	Moderate
Full Can (pain)	Torn Supraspinatus	Itoi[54]	80	50	Moderate
	RC Tear	Kim E[52]	71	32	Moderate
	Subacromial impingement	Kelly[40]	35	25	Low

Table 2 *Continued*

Test name(s)	Pathology	Lead Author	Sensitivity	Specificity	Risk of Bias* from QUADAS 2
Full Can (weak)	Torn Supraspinatus	Itoi[54]	83	53	Moderate
	RC Tear	Kim E[52]	77	32	Moderate
	Subacromial impingement	Kelly[40]	45	75	Low
Full Can (pain or weak)	Supraspinatus Tendinopathy	Chew[32]	75	68	Low
	RC Tear	Kim E[52]	90	54	Moderate
Full Can (pain and weak)	RC Tear	Kim E[52]	59	82	Moderate
External Rotation Lag Sign	Massive RC Tear	Jia[39]	35	89	High
	FTT- Supraspinatus	Bak[33]	45	91	High
	FTT- Supraspinatus	Castoldi[30]	56	98	Low
	FTT- Infraspinatus	Castoldi[30]	97	93	Low
	FTT- Teres Minor	Castoldi[30]	100	93	Low
	RC Tendinopathy	Jia[39]	7	84	High
	FTT- Supraspinatus/Infraspinatus	Miller[25]	46	94	Moderate
Hawkins-Kennedy	Supraspinatus Tendinopathy	Chew[32]	87	32	Low
	FTT- Supraspinatus	Bak[33]	77	26	High
	Subacromial impingement	Kelly[40]	74	50	Low
	Subacromial impingement	Michener[24]	63	62	Low
	Subacromial impingement	Silva[31]	74	40	Low
	Subacromial impingement	Fodor[27]	72	89	Moderate
	Subacromial impingement	Salaffi[43]	64	71	Moderate
	RC Tendinopathy	Fowler[37]	58	72	High
	AC Joint OA	Jia[39]	47	45	High
	Biceps Tendinopathy	Jia[39]	55	38	High
Internal Rotation Lag Sign	Subscapularis Tendinopathy	Bartsch[34]	71	60	Low
	FTT- Supraspinatus	Bak[33]	31	87	High
	Subscapularis Tear	Miller[25]	100	84	Moderate
Labral Tension	SLAP	Cook[50]	28	76	Moderate
Lateral Jobe	RC Tear	Gillooly[46]	81	89	Low
Lift-off	Partial Biceps Tear	Gill[51]	28	89	Low
	Biceps Tendinopathy	Jia[39]	28	89	High
	Subscapularis Tendinopathy	Bartsch[34]	40	79	Low
	Subscapularis Tendinopathy	Naredo[53]	50	84	Moderate
	Subscapularis Tendinopathy	Kim HA[42]	69	48	High
	Subscapularis Tendinopathy	Kim HA[41]	6	23	Low
	Subscapularis Tendinopathy	Salaffi[43]	35	75	Moderate
	Subacromial impingement	Silva[31]	68	50	Low
	RC Tendinopathy	Fowler[37]	19	90	High
	Glenohumeral OA	Jia[39]	29	90	High
	RC Tendinopathy	Jia[39]	10	79	High
Neer	Supraspinatus Tendinopathy	Chew[32]	64	61	Low
	FTT- Supraspinatus	Bak[33]	60	35	High
	Subacromial impingement	Kelly[40]	62	10	Low
	Subacromial impingement	Michener[24]	81	54	Low
	Subacromial impingement	Silva[31]	68	30	Low
	Subacromial impingement	Fodor[27]	54	95	Moderate
	AC Joint OA	Jia[39]	57	41	High
	Biceps Tendinopathy	Jia[39]	64	41	High
Olecranon Manubrium Percussion	Bony Abnormality	Adams[47]	84	99	Low
Painful Arc	Supraspinatus Tendinopathy	Chew[32]	71	81	Low
	Subacromial impingement	Kelly[40]	49	33	Low
	Subacromial impingement	Michener[24]	75	67	Low
	Subacromial impingement	Fodor[27]	67	80	Moderate
	FTT- Supraspinatus	Bak[33]	96	4	High
Palpation- biceps	Biceps Tendinopathy	Chen	57	74	Low
	Partial Tear- Biceps	Gill[51]	53	54	Low
	Type II SLAP	Oh[55]	27	66	Low
Palpation-coracoid	Adhesive Capsulitis	Carbone[48]	96	87	High
Passive-Abduction (pain)	Subacromial impingement	Silva[31]	74	10	Low
Passive Compression	SLAP	Kim YS[26]	82	86	Low
Passive Distraction	SLAP	Schlecter[45]	53	94	Low
Patte	Subacromial impingement	Silva[31]	58	60	Low
	Infraspinatus Tendinopathy	Kim HA[42]	63	73	High
	Infraspinatus Tendinopathy	Salaffi[43]	62	74	Moderate
	Infraspinatus Tendinopathy	Naredo[53]	71	90	Moderate
	Infraspinatus Tear	Naredo[53]	36	95	Moderate
Relocation	Type II SLAP	Oh[55]	44	54	Low
	Bankart lesion	Fowler[37]	79	87	High
	Hill-Sachs Lesion	Fowler[37]	81	81	High
Resisted- Abduction (pain)	Subacromial impingement	Kelly[40]	55	75	Low
Resisted-Abduction (weak)	Subacromial impingement	Kelly[40]	38	50	Low
	Subacromial impingement	Silva[31]	58	20	Low

Table 2 *Continued*

Test name(s)	Pathology	Lead Author	Sensitivity	Specificity	Risk of Bias* from QUADAS 2
Resisted-External Rotation/ Infraspinatus test (pain)	Subacromial impingement	Kelly[40]	33	90	Low
	Torn Infraspinatus	Itoi[54]	46	54	Moderate
	Infraspinatus Tendinopathy	Goyal[38]	50	100	High
	Subacromial impingement	Michener[24]	56	87	Low
Resisted-ER/Infraspinatus test (weak)	Subacromial impingement	Kelly[40]	55	25	Low
	Torn Infraspinatus	Itoi[54]	84	53	Moderate
Resisted-Lift-off (pain)	Torn Subscapularis	Itoi[54]	46	69	Moderate
Resisted-Lift-off (weak)	Torn Subscapularis	Itoi[54]	79	59	Moderate
Shoulder Shrug	Glenohumeral OA	Jia[28]	91	57	Low
	Adhesive Capsulitis	Jia[28]	95	50	Low
	RC Tendinopathy	Jia[39]	96	53	Low
	Massive RC Tear	Jia[28]	75	50	Low
Speed	SLAP	Cook[50]	28	76	Moderate
	Type II SLAP	Oh[55]	32	66	Low
	Labral Tear	Kibler[35]	29	69	Low
	SLAP	Ebinger[49]	60	38	Low
	Biceps Tendinopathy	Chen[36]	63	60	Low
	Partial Tear- Biceps	Gill[51]	50	67	Low
	Biceps Tendinopathy	Kibler[35]	54	81	Low
	Biceps Tendinopathy	Jia[39]	50	67	High
	Biceps Tendinopathy	Goyal[38]	71	85	High
	Biceps Tendinopathy	Salaffi[43]	49	76	Moderate
Supine Flexion Resistance	SLAP	Ebinger[49]	80	69	Low
Upper Cut	Biceps Tendinopathy	Kibler[35]	73	78	Low
	Labral Tear	Kibler[35]	22	56	Low
Yergason	Biceps Tendinopathy	Chen[36]	32	78	Low
	Biceps Tendinopathy	Kibler[35]	41	79	Low
	Biceps Tendinopathy	Kim HA[42]	14	89	High
	Biceps Tendinopathy	Kim HA[41]	75	81	Low
	Labral Tear	Kibler[35]	26	70	Low
	Type II SLAP	Oh[55]	12	87	Low
Yocum	Subacromial impingement	Silva[31]	79	40	Low
	Subacromial impingement	Fodor[27]	70	92	Moderate
Whipple	Type II SLAP	Oh[55]	65	42	Low
	RC Tendinopathy	Jia[39]	80	33	High
	Massive RC Tear	Jia[39]	100	26	High

*Bias: High= score of high risk of bias in 3 or 4 of total 4 categories; Moderate = score of high risk of bias in 2 of total 4 categories; Low = score of high risk of bias in 0 or 1 of total 4 categories. The 4 categories are: 1. Patient selection 2. Index test 3. Reference standard 4. Flow and timing.
AC, acromioclavicular; ER, external rotation; OA, osteoarthritis; RC, rotator cuff; SLAP, superior labrum anterior to posterior.

2. The greatest risk of bias was most often associated with the Q2 items Patient Flow and Reference Standard. The greatest concern in the category of applicability was also the reference standard with the addition of the index test. Patient flow concerns become apparent when there was an indeterminate or excessive time between the issuing of the index test and the criterion standard, when patients received different reference standards, or when it was difficult to discern if all patients were included in the analysis. Most of the studies, where patient flow was an issue failed to note the length of time between the index test and reference standard, or did not make clear whether all patients were included in the analysis. Often, there was an inability to reconstruct the 2x2 tables accurately from the data reported in the original article. The concern for bias in the reference standard was most often due to a failure to use a double blind design (issuer of the criterion standard was not blinded to index test result) or the failure to use the criterion standard to confirm diagnosis. The obvious gain in popularity of diagnostic ultrasound (n=12 studies in this review) had the deleterious effect of increasing concern for bias since ultrasound is not the criterion standard for shoulder diagnosis.[56–58] Lastly, the concern for applicability as it relates to the index test is because the authors failed to describe the index test.

Statistical analysis

Overall

Publication bias was not found to be evident with graphical or in statistical analysis. However, publication bias cannot be completely ruled out since these tests have decreased statistical power when analysing less than 10 studies.[59] No significant negative correlations were found to indicate the influence of threshold effects. Table 3 presents the results of meta-analysis for the individual ShPE tests by diagnosis, number of studies and sample size for the analyses.

Subacromial impingement

The Neer, Hawkins-Kennedy and painful arc tests for subacromial impingement were summarised for their diagnostic properties and associations. The strongest summary sensitivity was for the Hawkins-Kennedy test (0.80; 0.72, 0.86). However, the value was merely on the sensitivity threshold (80%) for assisting in ruling out subacromial impingement but because of poor specificity, the LR- value shows this test to have little effect on post-test probability to rule out subacromial impingement when negative. In fact, none of the three diagnostic tests demonstrated the likelihood ratios that would be unlikely to result in important changes in post-test probability. The pooled DOR

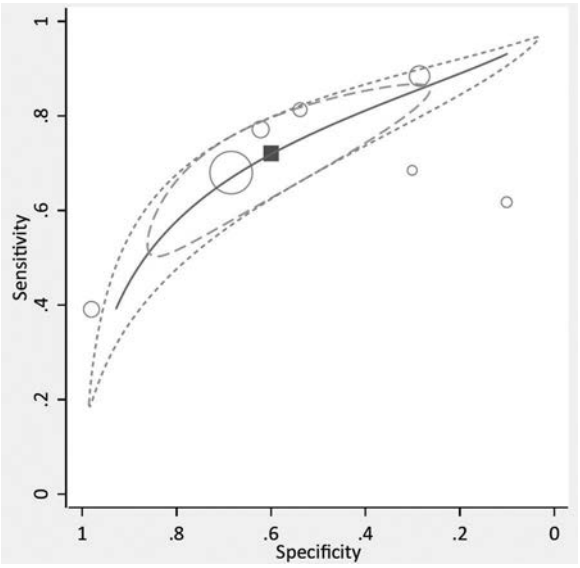

Figure 3 Hierarchical summary receiver operating characteristic (HSROC) curve composed of studies examining the diagnostic value of the Neer test in cases of subacromial impingement. This figure is only reproduced in colour in the online version.

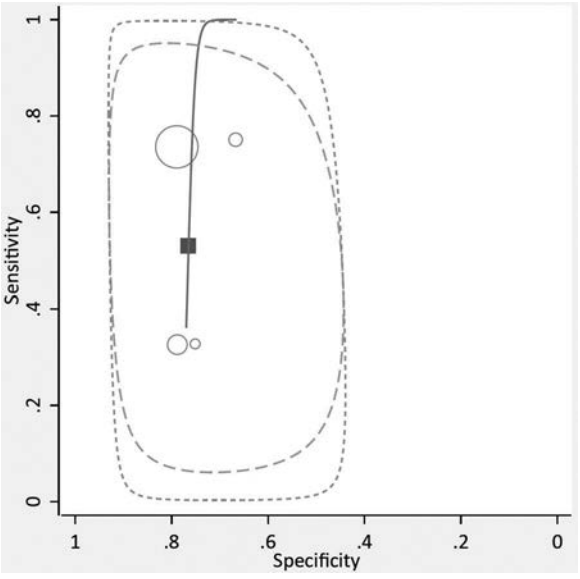

Figure 5 Hierarchical summary receiver operating characteristic (HSROC) curve composed of studies examining the diagnostic value of the Painful Arc test in cases of subacromial impingement. This figure is only reproduced in colour in the online version.

for any of these three tests indicates the discriminative diagnostic ability to determine a positive test result among those with subacromial impingement when compared with those without subacromial impingement is unlikely to occur. Figure 3 (Neer), figure 4 (Hawkins-Kennedy) and figure 5 (painful arc) illustrate the included studies with both the 95% confidence and prediction regions indicating the probable wide variability of the true sensitivity and specificity in future studies.

Meta-regression was conducted for both the Neer and Hawkins-Kennedy tests in order to determine if the summary DOR was biased as a result of differing reference standards. For the Neer test, there was a substantially greater DOR among the studies which used the gold standard of surgery for index test confirmation (4.85 ((95% CI 3.46 to 6.79)) than other reference

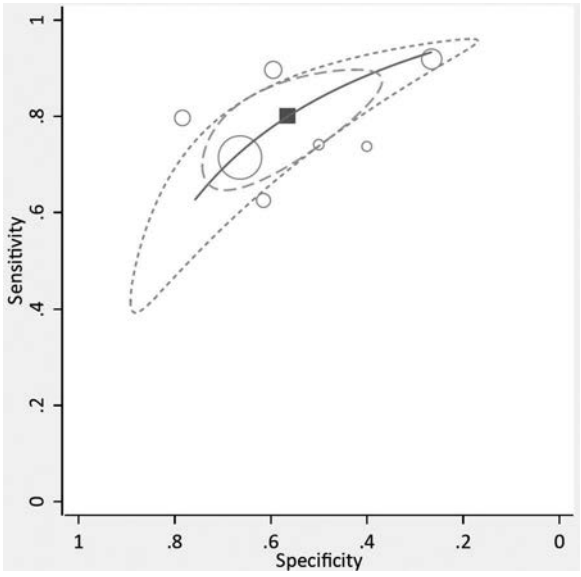

Figure 4 Hierarchical summary receiver operating characteristic (HSROC) curve composed of studies examining the diagnostic value of the Hawkins-Kennedy test in cases of subacromial impingement. This figure is only reproduced in colour in the online version.

standards (1.28 ((95% CI 0.31 to 5.19)). The ratio of DORs was strong (3.79 ((95% CI 0.87 to 16.14)) and the stratified estimates were statistically significant (p=0.07). Similarly, the DOR for the Hawkins-Kennedy test was stronger among those studies with the gold standard of surgery (6.41 ((95% 3.33 to 12.35) than for studies using other than the gold standard (3.14 ((95% 1.37 to 7.22)). However, the stratified estimates were not significantly (p=0.18) different from one another.

SLAP lesions

None of the 8 ShPE tests for which meta-analysis was possible (table 3) demonstrated sensitivity values that would likely rule out a SLAP lesion with a negative test. Yergason's test had the strongest summary specificity (95.3; 90.6,98.1), but again, the sensitivity was so poor that the LR+ demonstrates insignificant ability of this test to rule in a SLAP lesion when positive. All eight diagnostic tests for a SLAP lesion had likelihood ratios and DORs that were weak and their CI contained the null value (table 3).

The active compression test analysis found the O'Brien *et al*[60] study to have a large Cooks-D and standardised residuals influencing the summary estimates. Cooks-D is a measure of the influence that a particular study may have on the model parameters and can be used to check for particularly influential studies. Sensitivity analysis, with removal of the O'Brien *et al*[60] study, resulted in substantial attenuation of the DOR from 3.14 (95% CI 0.42 to 23.40) to 1.19 (95% CI 0.76 to 1.86). As such, this study was not included in summary estimates for the Active Compression test. Figure 6 illustrates the HSROC curves of the Active Compression test both with and without the outlier study.[60]

Anterior instability

Statistical pooling was done individually for three tests for the diagnosis of anterior instability: the apprehension, relocation and surprise tests. The surprise test demonstrated the strongest sensitivity (81.8; 69.1, 90.9), and therefore, negative likelihood ratio (0.25; 0.08–0.78)) that would likely rule out anterior instability when negative. All three tests demonstrated the ability to rule in anterior instability due to strong specificity.

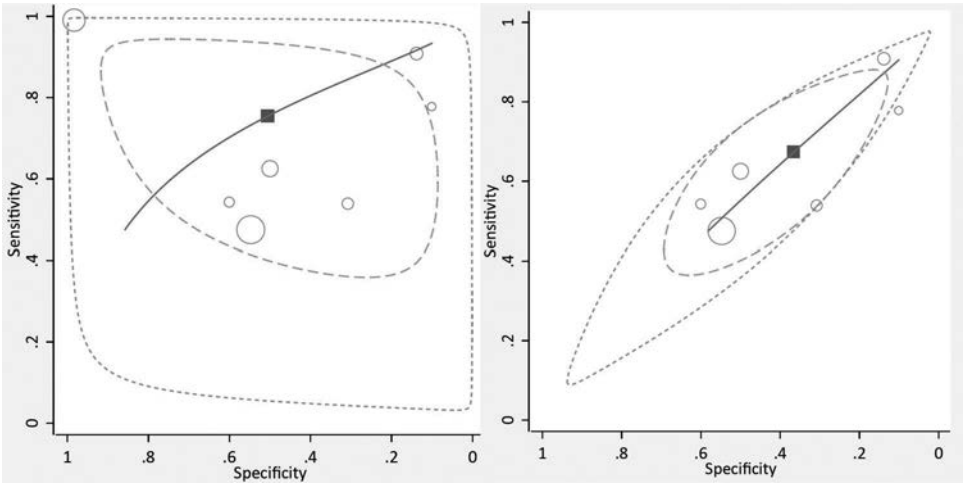

Figure 6 Hierarchical summary receiver operating characteristic (HSROC) curve composed of studies examining the diagnostic value of the Active Compression test in cases of a SLAP lesion. The left graph includes the original article reporting on the value of the test and the right graph shows the result of the elimination of this outlier study[60]. This figure is only reproduced in colour in the online version.

Table 3 Summary estimates from meta-analysis

Diagnosis Test	No. Studies Sample Size (n)	SN(95% CI)	SP(95% CI)	+LR(95% CI)	-LR(95% CI)	DOR(95% CI)
Impingement						
Neer*	7(n=946)	0.72(0.60, 0.81)	0.60(0.40, 0.77)	1.79(1.24, 2.58)	0.47(0.39, 0.56)	3.83(2.51, 5.84)
H-K*	7(n=944)	0.80(0.72, 0.86)	0.56(0.45, 0.67)	1.84(1.49, 2.26)	0.35(0.27, 0.46)	5.18(3.64, 7.35)
Painful Arc*	4(n=756)	0.53(0.31, 0.74)	0.76(0.68, 0.84)	2.25(1.24, 4.08)	0.62(0.37, 1.03)	3.66(1.24, 10.81)
SLAP						
Active Compression*	6(n=782)	0.67(0.51, 0.80)	0.37(0.22, 0.54)	1.06(0.90, 1.25)	0.89(0.67, 1.20)	1.19(0.76, 1.86)
Speeds*	4(n=327)	0.20(0.05, 0.53)	0.78(0.58, 0.90)	0.90(0.43, 1.90)	1.03(0.86, 1.23)	0.87(0.35, 2.55)
Anterior Slide*	4(n=831)	0.17 (0.03, 0.55)	0.86(0.81, 0.89)	1.20(0.22, 6.51)	0.97(0.96, 1.36)	1.24(0.16, 9.47)
Crank*†	4(n=282)	0.34(0.19, 0.53)	0.75(0.65, 0.83)	1.36(0.84, 2.21)	0.88(0.69, 1.12)	1.54(0.75, 3.18)
Yergason's	3(n=246)	12.4(6.60, 20.6)	95.3(90.6, 98.1)	2.49(0.97, 6.40)	0.91(0.84, 0.99)	2.67(0.99, 7.73)
Relocation	3(n=246)	51.6(41.2, 61.8)	52.4(44.0, 60.6)	1.13(0.88, 1.45)	0.93(0.72, 1.20)	1.23(0.72, 2.11)
Biceps Palpation	2(n=114)	38.6(26.0, 52.4)	66.7(52.9, 78.6)	1.06(0.66, 1.68)	0.95(0.74, 1.22)	1.13(0.51, 2.50)
Compression Rotation†	2(n=355)	24.5(13.8, 38.3)	78.0(72.9, 82.5)	2.81(0.20, 39.70)	0.87(0.66, 1.16)	3.39(0.15, 74.78)
Anterior Instability						
Relocation†	3(n=509)	64.6(54.9, 73.4)	90.2(86.8, 93.0)	5.48(0.56, 53.8)	0.55(0.24, 1.27)	10.64(0.32, 354.10)
Apprehension	2(n=409)	65.6(52.7, 77.1)	95.4(93.3, 97.8)	17.21(10.02, 29.55)	0.39(0.22, 0.68)†	53.60(24.29, 118.30)
Surprise	2 (n=128)	81.8(69.1, 90.9)	86.1(72.1, 94.7)	5.42(0.96, 30.52)†	0.25(0.08, 0.78)†	28.10(7.71, 102.45)
Tendinopathy						
H-K	3(n=738)	65.5(60.3, 70.5)	62.8(57.3, 68.1)	1.86(1.47, 2.34)	0.46(0.36, 0.60)	4.68(3.35, 6.53)
Labral Tear						
Crank	3(n=187)	57.3(47.2, 67.0)	72.6(61.8, 81.8)	2.44(0.69, 8.59)	0.51(0.21, 1.22)	5.81(0.47, 71.50)

SN= sensitivity, SP=specificity, +LR=positive likelihood ratio, -LR=negative likelihood ratio, DOR=diagnostic odds ratio, CI=confidence interval, SLAP=……., *HSROC/Bivariate models and all others use DerSimoninian-Laird random-effects models. †indicates those studies and properties demonstrating significant heterogeneity (p>0.10).

The apprehension test had the strongest positive likelihood ratio (17.2; 10.02, 29.55) and was the only one of the three in which the CI did not contain the null value. The apprehension test had the strongest DOR (53.6; 24.3, 118.3), indicating some evidence for this test's overall diagnostic discriminative performance.

Significant heterogeneity was found in the properties and associations for the relocation test. Subgroup analysis, accomplished by removing the study by Lo *et al*[61] based upon the non-criterion reference standard used, did not improve the overall heterogeneity.

Labral tear

In pooled analyses, the crank test for labral tear demonstrated limited ability to rule in a labral tear with a +LR of 2.4 and specificity of 76%, indicating a likely small change in post-test probability.

Tendinopathy

In pooled analyses, the Hawkins-Kennedy test for tendinopathy demonstrated no evidence for the ability to rule in or out, change post-test probability or have overall diagnostic discriminative performance.

What this study adds

- This is the first meta-analysis to study ShPE tests and use the QUADAS 2 document to assist in the qualitative review and the HSROC/bivariate models for meta-analysis
- There is less optimism that the biceps load II is diagnostic for SLAP lesions
- The belly-off and modified belly press tests may be helpful in diagnosing subscapularis tendinopathy
- The bony apprehension test may help diagnose bony instability
- The olecranon-manubrium percussion test may be useful in a traumatic injury for bony abnormality requiring referral for x-ray
- The passive compression test may be helpful in diagnosing a SLAP lesion
- The modified dynamic labral shear test may be diagnostic of labral tears
- The lateral Jobe test may be useful for diagnosing a rotator cuff tear
- The shrug sign appears to be a sensitive test for stiffness-related disorders (osteoarthritis and adhesive capsulitis) as well as rotator cuff tendinopathy
- The passive distraction test may be able to rule in a SLAP tear if positive

DISCUSSION

This is the first study on diagnostic accuracy of which we know that has incorporated HSROC/bivariate models as the criterion standard during performance of a meta-analysis of ShPE tests. We feel that the use of this criterion standard promotes increased attention on and isolation of studies that demonstrate results dramatically outside others of similar context. Of particular interest, is the dramatic change in both the 95% CI and 95% prediction region of the active compression test for a SLAP lesion when the original study[60] is eliminated (figure 6). Further, this study[60] is a good example of the effect of bias on estimates of diagnostic accuracy given that the publication possesses examples of at least seven kinds of bias. Most notable of these biases, is partial verification bias which has been shown to overestimate the diagnostic accuracy of a test.[62]

For each diagnostic category, the overall results of this systematic review and meta-analysis indicate that a few tests are helpful to confirm or screen for a given diagnosis. There is a preponderance of evidence about individual physical examination tests that could not be combined for the meta-analysis. For those tests, we have used the diagnostic values and risk of bias from the Q2 to determine which tests are recommended for use as a screen or those recommended as a confirmatory test using the benchmarks of specificity >80%, sensitivity >80%, LR+ ≥ 5.0 and LR– ≤0.20. The list is short, and confidence in the diagnostic accuracy estimates is tenuous.

From the meta-analysis portion of this review, the Hawkins-Kennedy initially appears to be of value in ruling out subacromial impingement when negative. However, the LR– is poor and further, a strong argument can be made that subacromial impingement is not a valuable diagnosis but rather a cluster of diagnoses.[63] The diagnosis of subacromial impingement encompasses such a broad range of pathologies, from bursitis to a complete rotator cuff tear,[64] that a label of subacromial impingement may not help guide treatment.[65] Yergason's test, used for detection of a SLAP lesion, has high (95%) pooled specificity. However, the sensitivity is so low, that a positive test modifies the post-test probability of detecting a SLAP lesion only a small amount. In a similar perspective to subacromial impingement, authors have argued that tests results for SLAP may be effected by the percentage of different forms of Snyder classifications present within the sample.[50]

Therefore, the only tests that appear to have good clinical utility are the apprehension, relocation, and surprise tests to diagnose anterior instability and these tests are primarily a continuum of the apprehension test. When a patient registers apprehension with this test, the relocation manoeuvre should then decrease apprehension, whereupon, the relocation force is removed causing a surprised reaction (surprise test) by the patient as the apprehension reappears.

While the results of the meta-analysis were, perhaps, not inspiring to the clinician searching for diagnostic answers, there are some individual tests that warrant further investigation.

Table 4 Best* Test Combinations and Reported Value for Various Pathologies

Test Combination	Pathology	Lead Author	Sensitivity	Specificity	Positive LR	Negative LR
Passive Distraction and Active Compression	SLAP	Schlecter[45]	70	90	7.00	.11
Compression-rotation AND Apprehension AND Speed	Type II SLAP	Oh[55]	25	92	3.13	0.82
Anterior Slide AND Crank	Labral Tear	Walsworth[44]	34	91	3.75	0.73
Apprehension AND Relocation	Labral Tear	Guanche[66]	38	93	5.43	0.67
Age>39, Painful Arc, Self-report of Popping or Clicking	Supraspinatus Tendinopthy	Chew[32] ≥ 2 positive tests; 3 positive tests	75, 38	81, 99	3.82, 32.20	0.32, 0.63
Age≥65 AND Weakness in ER (Infraspinatus Test) AND Night Pain	RC Tear	Litaker[67]	49	95	9.84	0.54
Hawkins-Kennedy, Neer, Painful Arc, Empty Can, Resisted ER	Subacromial impingement	Michener[24]; ≥ 3 positive tests	75	74	2.93	0.34
Lift-off and/or Resisted IR	Subscapularis Tendinopathy; Subscapularis Tear	Naredo[53]; Naredo[53]	50, 50	84, 95	3.13, 10.0	0.60, 0.53
Apprehension AND Relocation	Anterior Instability	Farber[68]	81	98	39.68	0.19

*Best is defined as the highest sensitivity, specificity, or both from the studies with the least bias.

The posterior apprehension test for posterior instability demonstrated a higher specificity and positive likelihood ratio but these values came from a high bias study.[39] Another highly specific test, but from a low bias study[45] is the passive distraction test for a SLAP lesion. This test may rule in a SLAP lesion when positive. Sensitive tests of note are the shoulder shrug sign, for stiffness-related disorders (osteoarthritis and adhesive capsulitis) as well as rotator cuff tendinopathy and the Whipple test for massive rotator cuff tears. However, the diagnostic properties of the Whipple test come from a high bias study.[39] Other tests of possible value from high bias studies included the AC resisted extension,[39] the resisted belly press,[38] and coracoid palpation.[48] There are six additional tests with higher sensitivities, specificities, or both but caution is urged since all of these tests have been studied only once and more than one ShPE test (ie, active compression, biceps load II) has been introduced with great diagnostic statistics only to have further research fail to replicate the results of the original authors. The belly-off and modified belly press tests for subscapularis tendinopathy, bony apprehension test for bony instability, olecranon-manubrium percussion test for bony abnormality, passive compression for a SLAP lesion, and the lateral Jobe test for rotator cuff tear give reason for optimism since they demonstrated both high sensitivities and specificities reported in low bias studies. Finally, one additional test was studied in two separate papers.[35 50] The modified dynamic labral shear test, may be diagnostic of labral tears in general, but be sensitive for SLAP lesions specifically.

Looking back to our initial publication and combining that data with the current review certainly expands the clinician's diagnostic arsenal. The external rotation lag sign continues to be recommended as it was in 2008[1] to confirm full-thickness rotator cuff tears of the infraspinatus. The hornblower's sign may be diagnostic of severe degeneration or absence of the teres minor muscle, and the active compression test may have value as a confirmatory test for AC joint pathology when positive due to its high specificity.

Despite some cause for optimism when looking at some of the individual studies and tests, the more prudent method may be to look at clusters or combinations of tests, since that resembles more closely, the way in which most ShPE tests are used in the clinic. Table 4, while not all-inclusive, shows the best test combinations to date for detecting various pathologies.

Unfortunately, even many of these test clusters modify the post-test probability by a small to minimal amount. Of note in this group of clustered tests is the combination of age>39, painful arc, and self-report of popping and clicking[32] and the combination of the apprehension and relocation tests,[68] both of which produce a large post-test shift toward the diagnoses of supraspinatus tendinopathy, and anterior instability, respectively.

LIMITATIONS

Any review is limited by the quality of studies contained therein. Many of the studies in this review had issues with the reference standard and subject flow and timing. There was clearly a rise in the use of diagnostic ultrasound as a criterion standard, and evidence to supports its use is currently poor.[56–58] Further, we limited our articles to those in the English language which may make this review more prone to dissemination bias. However, publication bias was not found to be evident with graphical or in statistical analysis. Finally, this is the first meta-analysis on diagnostic accuracy of ShPE tests that was performed using the Q2 document. The original authors piloted the Q2 on five studies and found that reliability varied considerably.[14] Our weighted κ (κ=0.31; 0.10, 0.52) was likewise only fair.

CONCLUSIONS

Based on data from our original review[1] and this update, the use of any single ShPE test to make a pathognomonic diagnosis cannot be unequivocally endorsed due to continued quality issues in publications. Some ShPE tests are beginning to stand the tests of scrutiny and time but there are far more tests that need to be validated in more than one study. Combinations of ShPE tests provide better accuracy, but marginally so. These findings seem to provide support for stressing a comprehensive clinical examination including history and clinical examination. However, there is a great need for large, prospective, well-designed studies that examine the diagnostic accuracy of the many aspects of the clinical examination and what combinations of these aspects are useful in differentially diagnosing pathologies of the shoulder.

Acknowledgements The authors would like to acknowledge Ms Connie Schardt for her invaluable assistance in the search process and the authors from the original paper whose initial work was foundational:S Campbell, A Morin, M Tamaddoni, C T Moorman III.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

► References to this paper are available online at http://bjsm.bmjgroup.com