Hegedus EJ, Goode A et al. Physical examination tests of the shoulder: a systematic review with meta-analysis of individual tests. Br J Sports Med 2008;42:80-91.

Design: Meta-analysis of diagnostic tests

Population/sample size:

- 45 articles evaluating physical examination tests of the shoulder, selected from 922 electronically retrieved abstracts and hand searches
- Inclusion criteria were: gold standard was surgery, MRI, or injection (AC joint only); at least one physical exam test studied, sensitivity & specificity reported or discernible from reported results, published in English
- Excluded if test done under anesthesia or on cadavers, if a group of tests was reported as a composite physical exam, or if article was a review
- Quality Assessment of Diagnostic Accuracy Tests (QADAS) was used to score individual articles; QADAS has scale from 0 to 14; with scores of 10 or more considered high quality

Main outcome measures:

- Meta-analysis could be done on only three tests where there was sufficient homogeneity to draw summary conclusions: the Neer test for impingement, the Hawkins test for impingement, and the Speed test for SLAP lesions
- For the Neer test, the pooled sensitivity was 79% and the specificity was 53%; the 95% confidence interval for the pooled diagnostic odds ratio crossed 1, meaning that it was not significant and has no diagnostic utility
- For the Hawkins test, the pooled sensitivity was 79% with specificity of 59%; as with the Neer test, the diagnostic odds ratio was not significant
- For the Speed test, the pooled sensitivity was 32% with specificity of 61%; the diagnostic odds ratio also was not significant
- For rotator cuff pathology, none of the 10 physical tests examined in more than one study proved consistently diagnostic
- For a tear of any rotator cuff muscle, the External Rotation Lag Sign (ERLS) and the drop arm test were specific enough to confirm the diagnosis
- The supine impingement test may be sensitive enough to rule out a rotator cuff tear
- For impingement without a tear, no test is sensitive enough to rule out a tear, but the empty can test and the infraspinatus tests may be specific enough to confirm impingement; this conclusion should be regarded with caution, since only one study examined their accuracy
- The bear-hug and belly press tests appear to be specific enough to rule in a subscapularis tear when positive
- For glenoid labrum integrity, the biceps load II test appears diagnostic for SLAP but further study is needed
- For instability, meta-analysis was not possible, but the apprehension, relocation, and anterior release tests appear diagnostic, provided that apprehension and not pain is used as the end point of the test

- For AC joint pathology, no meta-analysis was possible, but the O'Brien test may be diagnostic; however, the higher quality studies reported less diagnostic accuracy than the lower quality studies
- For AC joint pathology, pain with palpation may be sensitive enough to serve as a screening test
- For cuff tears, the hornblower's sign may be diagnostic of severe degeneration of teres minor

Authors' conclusions:

- Very few physical examination signs appear to be diagnostically discriminatory and most are not useful in the clinic
- Only two studies were sufficiently powered to detect a physical sign with high sensitivity; all conclusions must be drawn with caution
- Large and methodologically robust studies of physical exam tests are needed

Comments:

- Well-established (QADAS) quality criteria have been applied, but a study with a high score may still have limitations
- For example, the biceps II study had a high QADAS score, but the paper itself did not make it clear that the surgeons who decided on the gold standard were unaware of the results of the physical exam; they may themselves have been the examiners
- Figures 2, 3, and 6 appear to be incorrect in several respects; they are labeled as diagnostic odds ratios, but since there cannot be negative odds ratios, it is clear that they must refer to the natural logarithm of the odds ratios
- In addition, the Cochrane Handbook (Bossuyt 2013, page 11) states that diagnostic odds ratios should not be considered a suitable summary test statistic to describe test performance
- Even with this stipulation, figures and 3 are suspect: the four individual studies have confidence intervals that are statistically significant (log odds does not cross 0; therefore odds ratio does not cross 1), but the pooled odds ratio does cross the value for 1; this was reported by the authors as showing that the Neer and Hawkins tests were non-discriminatory
- Review of the four source papers for figures 2 and 3 was performed by calculating the numbers of patients in each of the articles with impingement and the numbers with positive and negative tests; when pooled, the pooled answer is approximately that of the authors for figure 2 (a pooled log odds of 1.54 is a diagnostic odds ratio of 4.66 by the authors; my answer is 1.59 with an odds ratio of 4.9)
- However, the pooled odds ratio does not cross 1; the 95% confidence intervals are between 3.5 and 6.8 using SPSS software (attached)
- Even though there appears to be an error in the pooling, and the diagnostic odds ratio probably is significantly greater than 1, the high end of the 95% confidence interval (6.8) is still too low to represent a useful test; diagnostic odds ratios of 20 or more are generally considered to be acceptable

- Some of the authors' endorsements (biceps II, supine impingement, hornblower's, bear hug, belly press) are based on single studies, and their caveats about these results should be emphasized
- In spite of analytical difficulties, the basic conclusions concerning the limited discriminatory power of single physical exam tests appear reasonable
- Combinations of physical exam tests were not evaluated; the conclusion that single tests are not diagnostic is basically uncontroversial

Assessment: Inadequate due to incorrect analysis of data; when calculations are redone, the basic conclusion (that the Neer, Hawkins, and Speed tests are not discriminatory enough to rule in or rule out shoulder diagnoses) still appears justified

Reference:

Bossuyt P, Davenport C, et al. Chapter 11:Interpreting results and drawing conclusions. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9. The Cochrane Collaboration, 2013. Available from: http://srdta.cochrane.org/.