Physical Therapy in Sport xxx (2014) 1–6

Contents lists available at ScienceDirect

Physical Therapy in Sport

journal homepage: www.elsevier.com/ptsp

Masterclass

Combining orthopedic special tests to improve diagnosis of shoulder pathology

Eric J. Hegedus [a, *], Chad Cook [b], Jeremy Lewis [c], Alexis Wright [a], Jin-Young Park [d]

[a] High Point University, Department of Physical Therapy, High Point, NC 27262, USA
[b] Physical Therapy Program, Duke University, Durham, NC, USA
[c] Physiotherapy, University of Hertfordshire, Department of Allied Health Professions and Midwifery, School of Health and Social Work, United Kingdom
[d] Shoulder, Elbow & Sports Center, Konkuk University, Seoul, South Korea

ARTICLE INFO

Article history:
Received 2 April 2014
Received in revised form
5 July 2014
Accepted 1 August 2014

Keywords:
Likelihood ratios
Shoulder
Diagnosis

ABSTRACT

The use of orthopedic special tests (OSTs) to diagnose shoulder pathology via the clinical examination is standard in clinical practice. There is a great deal of research on special tests but much of the research is of a lower quality implying that the metrics from that research, sensitivity, specificity, and likelihood ratios, is likely to vary greatly in the hands of different clinicians and in varying practice environments. A way to improve the clinical diagnostic process is to cluster OSTs and to use these clusters to either rule in or out different pathologies. The aim of the article is to review the best OST clusters, examine the methodology by which they were derived, and illustrate, with a case study, the use of these OST clusters to arrive at a pathology-based diagnosis.

1. Introduction

Physical examination of the shoulder involves a series of steps typically beginning with history, progressing with motion and muscle testing, and culminating in the use of orthopedic special tests (OSTs) with the aim of diagnosing shoulder pathology. While the process itself is systematic and straightforward, for evidence-based practitioners, there are numerous problems encountered when trying to arrive at a diagnosis. First, there is little evidence reporting the diagnostic accuracy of critical pieces of the clinical examination such as history, motion testing, and muscle testing causing a greater reliance on OSTs. Second, although there is a great deal of research on OSTs of the shoulder, much of that research is of moderate to low quality (Hegedus et al., 2008; Hegedus et al., 2012). Third, even in those OSTs that come from high quality literature, there are very few that display solid diagnostic metrics, high sensitivity and specificity (Hegedus et al., 2008; Hegedus et al., 2012). Fourth, although sensitivity and specificity are helpful internal test metrics, there are issues in the application of these metrics to clinical practice. Finally, clinicians and researchers improve diagnostic accuracy by clustering OSTs together; however, in some cases, the clusters are used incorrectly or provide metrics that lead to post-test probabilities that are no different than use of a single stand alone test.

Our aims in this paper are to discuss the importance of likelihood ratios and modified probability in the diagnostic process, to explain multivariate modeling and outline the most effective methods to combine tests for either screening or confirmation of diagnosis. For context, we'll briefly review the best test clusters that have been published, and finally, we'll use a case study to illustrate how the best available test clusters should be used to aid in diagnosis.

2. Likelihood ratios and modified probability

Diagnostic accuracy studies have design consistencies, standardized metrics, and assumptions. First and foremost, all diagnostic accuracy studies enroll populations of individuals with and without the condition of interest; the condition of interest being the diagnosis studied. Those without the condition of interest should be individuals with some other competing health malady that would normally be distinguished in a traditional clinical environment. For example, a typical shoulder study enrolls patients with pain, some of whom will have the condition of interest, like a rotator cuff tear, and some of whom will have other usual sources of shoulder pain like tendinosis or a labral tear.

The simplest measures of diagnostic accuracy are sensitivity and specificity. Sensitivity is the proportion of people *with the condition*

* Corresponding author. Tel.: +1 336 906 2133.
E-mail address: ehegedus@highpoint.edu (E.J. Hegedus).

http://dx.doi.org/10.1016/j.ptsp.2014.08.001
1466-853X/

of interest who will have a *positive* result, whereas specificity is the proportion of the patients who *do not have the condition of interest* who have a *negative* result (Table 1). Mathematically, sensitivity values are calculated only from those with the condition of interest, whereas specificity values are calculated from those without the condition of interest. This is one reason that the use of these internal metrics is limited. For example, sensitivity fails to recognize any of the examination findings that are reflective of those who did not have the condition of interest.

Both sensitivity and specificity are reported in percentages, from 0 to 100. A 100% sensitivity or specificity suggests that the test will be positive 100% of the time (if truly sensitive) in patients with the condition of interest and will be appropriately negative in 100% of cases (if truly specific) when the patient does not have the condition of interest. In order to emphasize the context associated with these measures, a sensitivity of 20%, which is a finding associated with most reflex testing, suggests that the test will be positive in only 20% of cases in which the patient actually has the condition of interest, often a neurological problem like radiculopathy.

Likelihood ratios and probability metrics are calculated from the sensitivity and specificity values (Table 1). A positive likelihood ratio (LR+) is derived from subjects with and without the condition of interest. LR+ is calculated by taking the sensitivity and dividing by 1-specificity. With LR+, positive findings influence the post-test decision-making and a stronger LR+ will be notably greater than 1.0. In contrast, negative likelihood ratio (LR−) is calculated by taking 1-sensitivity divided by the specificity, where a robust finding is hallmarked by smaller values closer to 0, and reflects a negative test finding only. Both LR− and LR+ are used to calculate post-test probabilities (0–100%) when the tests are negative or positive. A strong clinical test (or cluster of tests) should have the ability to rule out a condition when negative (with post-test probabilities near 0) or rule in a condition when positive (with post-test probabilities near 100).

Further, truly robust tests should have confidence intervals that are precise, which is suggestive that repeating the study findings should lead to similar results. A confidence interval is a parameter estimate that outlines the boundaries that a given test value will fall within if the study is repeated numerous times. In most cases, 95% confidence intervals are reported. This means that there is a 95% chance that the test value would fall within the boundaries of the confidence interval if the study was replicated in a different sample. For example, if an LR+ for a new labral tear test was 10.0 and the 95% confidence interval reported was 9.2–10.4, this means that there is a 95% chance that another trial would again find an LR+ between the boundaries of 9.2–10.4. These more precise confidence intervals would lead the reader to have greater confidence in the reported LR+ than if a wide confidence interval was reported, for example, 10.0 (1.0, 32.0). Remember that a likelihood ratio of 1.0 is not valuable and in this example, there is a chance that the true likelihood ratio could be 1.0.

Table 1
Sensitivity, specificity, and likelihood ratios.

↓Clinical test I reference test →	Positive test (often surgical confirmation)	Negative test (negative surgical findings)
Positive test (often pain or weakness)	True positive (*A*)	False positive (*B*)
Negative test (no pain or weakness)	False negative (*C*)	True negative (*D*)

Formulas:
Sensitivity $= A/(A + C)$.
Specificity $= D/(B + D)$.
Positive Likelihood Ratio (LR+) = Sensitivity/(1 − Specificity).
Negative Likelihood ratio (LR−) = (1 − Sensitivity)/Specificity.

Because LR+ and LR− determine the values of both positive and negative findings, evaluate individuals with and without the condition of interest, and can be used to estimate post-test probabilities with adjustments for pre-test prevalence, these metrics should be used to guide decision making over sensitivity and specificity. Both LR+ and LR− are calculated using both those with and without the condition of interest and only these measures truly reflect a situation of diagnostic uncertainly.

3. Multivariate modeling

The goal in any data analysis is to extract from raw information the accurate estimation (Alexopoulos, 2010). The goal when clustering tests is to determine the best combination estimates that produce the strongest likelihood ratios and to do so, multivariate modeling is required. Thus, clustering is simply the act of evaluating a set of tests and measures, in combination, when making a clinical decision or a mathematical assessment. For example, when attempting to detect acromiocalvicular (AC) joint pathology, two or more positive findings of cross-body adduction, the AC resisted extension test, and the active compression test, have an LR+ of 7.36 and an LR− of 0.21 which are better metrics than, for example, the active compression test alone with an LR+ of 1.6 and an LR− of 0.93 (Chronopoulos, Kim, Park, Ashenbrenner, & McFarland, 2004; Walton et al., 2004). Clustering tests more closely reflects how many clinicians make decisions because it takes into account a number of findings from the clinical assessment.

Multivariate modeling is a form of statistical analysis that explores the relationship between two or more predictor variables (the clinical tests) and the outcome variable (the reference standard). There are multiple forms of multivariate modeling methods and for clustering best tests and measures for diagnosis, a logistic regression analysis is the most appropriate type since the diagnosis is almost always dichotomous (present or absent).

Following proper multivariate modeling methodology is essential and the failure to do so when developing clustered rules or guides has been recognized by many authors (Beattie & Nelson, 2006; Beneciuk, Bishop, & George, 2009; Cook, Shah, & Pietrobon, 2008; Haskins, Rivett, & Osmotherly, 2012; May & Rosedale, 2009; Nee & Coppieters, 2011; Stanton, Hancock, Maher, & Koes, 2010). Although multivariate modeling can be notably complex, before considering clustering tests to determine most parsimonious values, it is useful to contemplate the following four principles: 1) determination of observations per variable, 2) linearity continuous measures, 3) assessment of conditional dependence of the predictor variables (also recognized as Variance Inflation Factor or Tolerance) and 4) appropriate stepwise modeling. The following paragraphs will provide recommendations for each of these principles.

3.1. *Determination of observations per variable*

There are a number of ways to determine the appropriate observations per variable, or, in determining how many tests should be included as independent variables in the multivariate model. For the sake of clarity, an observation would be an individual who is enrolled in the sample for the diagnostic accuracy study. For simple univariate, multinomial, or logistic regression, Hosmer and Lemeshow (2000) have recommended a minimum observation-to-variable ratio of 10, but cautioned that a number this low will likely overfit (overly burden) a model. That said, these same authors recommended the preferred observation-to-variable ratio of 20 to 1 for stepwise, multivariate modeling. A greater than 20:1 ratio is likely to provide more precise results. Thus, if one wanted to include 4 tests in a cluster multivariate model, a minimum total of a least 80 patients are recommended.

Please cite this article in press as: Hegedus, E. J., et al., Combining orthopedic special tests to improve diagnosis of shoulder pathology, *Physical Therapy in Sport* (2014), http://dx.doi.org/10.1016/j.ptsp.2014.08.001

3.2. Linearity of continuous measures

In most cases, clinical tests of the shoulder are either "positive" or "negative". In some cases, a positive finding is determined after a threshold score is ascertained from a continuous set of measures (e.g., a diagnosis of rotator cuff tear will only be made if active abduction is less the 90° threshold). When underlying tests have a continuous value (Like active or passive motion), the linearity of that value must be evaluated prior to determining a threshold. Linearity is generally analyzed by plotting to identify potential curvilinear relationships. An example of a lack of linearity is the estimation of the relationship of one's flexibility to injury. It is suggested that those who are overly inflexible and those who are excessively flexible, are more predisposed to an overuse injury that one who is in the middle ranges. If we were to evaluate the influence of flexibility toward injury on values such as these, values that are not linear, no significant relationship would occur. There are adjustments one can make if a variable lacks linearity. One can create categories and enter the variable as ordinal data with a set of indicators (dummies). Or, one can modify the definition of a positive test to reflect the variability within the underlying data. Using our previous flexibility example, one could score inflexible and excessively flexible as 'positive' and those in the middle categories as 'negative'.

3.3. Assessment of conditional dependence of the predictor variables

One possible reason why past studies have failed to outline clusters of findings is the concept of conditional dependence. Conditional dependence (Menten, Boelaert, & Lesaffre, 2008) occurs when a subsequent test finding is not dissimilar to the first test finding or when a series of tests actually measure the same thing and are positive together in clusters or negative together in clusters. During multivariate modeling, this dependence is routinely referred to as assessment of multicollinearity. One can assess multicollinearity through use of correlation matrixes, variance inflation factors (VIF) and tolerance values. A correlational finding of $r > 0.7$ between test variables can be used to assess the potential of multicollinearity. (Shen & Gao, 2008) A mean VIF close to 1 represents little collinearity, whereas 10 or greater is very poor and reflects very high collinearity. (Kutner, Nachtsheim, & Neter, 2004) Tolerance is the reciprocal of VIF thus values close to 0 are considered to have high collinearity. (Firth, 1993) If tests are conditionally dependent and are included in the multivariate model, there is a risk that the test will be removed from the final model and/or the test will remain in the model but will adversely influence the beta scores of the variables within the model. In layman's terms, a beta score allows for a consistent and meaningful measure across different units for the relationship of an independent variable to the dependent variable. A notable example of conditional dependence was the recent publication on clustered tests for cervical myelopathy by Cook, Brown, Isaacs, Roman, Davis, and Richardson (2010). In this study, various forms of reflex testing were nearly always hyper-responsive at different areas (quadriceps, Achilles, brachioradialis, etc.), but only one of the tests (positive brachioradialis) was included in the clustered model. Adding all of the reflex findings to the final model would not improve the accuracy and could alter the beta estimates.

3.4. Appropriate stepwise modeling

Stepwise modeling is not without controversy. (Wlikinson & Dallal, 1981) Stepwise regression modeling is an automatic procedure in which the choice of predictive variables is carried out until the strongest, most refined explanatory model is determined. Most commonly, univariate analyses for each single test to the reference standard is calculated. Univariate analyses with p values of <0.15 are generally included in a multivariate model since the tests interactions with other tests may yield diagnostic findings in the final analyses. Using the automated stepwise processes, most statistical software programs will calculate a single cluster of independent variables (tests) that are responsible for the best explanation. By analyzing the best number of positive findings (e.g., 1 of 4, 2 of 4, 3 of 4, etc) one can further determine the desired sensitivity and specificity of their created cluster.

4. Best test clusters

Before presenting the best published test clusters, "best" needs to be put in context. "Best" as used in this manuscript, is defined as those combinations of tests with the strongest likelihood ratios from research with the highest quality. The quality of the tests clusters is judged by using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) document and using a 0–14 (0 = lowest quality) scale (Whiting, Rutjes, Dinnes, Reitsma, Bossuyt, & Kleijnen, 2004). From our past experience(Cook & Hegedus, 2013; Hegedus et al., 2008), those studies scoring below 10/14 are full of design faults that make the likelihood ratios from those studies questionable and probably not repeatable in populations outside of those examined in the study.

Only 4 articles (Farber, Castillo, Clough, Bahk, & McFarland, 2006; Guanche & Jones, 2003; Litaker, Pioro, El Bilbeisi, & Brems, 2000; Park, Yokota, Gill, El Rassi, & McFarland, 2005) met our quality criteria and these articles reported on just 6 current clusters. The best test clusters currently available are summarized in Table 2. Unfortunately, even these high quality studies have failed in some respect with regard to sample size, stepwise regression, conditional dependence, and linearity of continuous measures (Table 3). Closer examination of these 4 studies reveals some other interesting findings. With regard to rotator cuff tears, of note is that both test clusters (Litaker et al., 2000; Park et al., 2005) incorporate older age as a component and that the most diagnostic cluster (Park et al., 2005) uses 2 tests, painful arc (Bak et al., 2010; Litaker et al., 2000; Michener, Walsworth, Doukas, & Murphy, 2009) and drop arm (Bak et al., 2010), that have low specificity or sensitivity values resulting in likelihood ratios that approach 1.0. In addition, another set of the tests from this same study (Park et al., 2005), the infraspinatus and painful arc tests, are part of the diagnostic cluster for

Table 2
Best test clusters from current literature.

Author(s)	Pathology	Test cluster	LR+	LR−
(Litaker et al., 2000)	Rotator cuff tear	1 Age > 65 and 2 Weakness in external rotation and 3 Night pain	9.84	0.54
(Park et al., 2005)	Rotator cuff tear (full thickness)	1 Age ≥ 60 and 2 + painful arc test and 3 + drop arm test and 4 + infraspinatus test	28.0	0.09
(Park et al., 2005)	Impingement	1 + Hawkins–Kennedy and 2 + painful arc test and 3 + infraspinatus test	10.56	0.17
(Farber et al., 2006)	Anterior instability (traumatic)	1 + apprehension test and 2 + relocation test	39.68	0.19
(Guanche & Jones, 2003)	Labral tear	1 + relocation test and 2 + active compression test	4.56	0.65
(Guanche & Jones, 2003)	Labral tear	1 + relocation test and 2 + apprehension test	5.43	0.67

Please cite this article in press as: Hegedus, E. J., et al., Combining orthopedic special tests to improve diagnosis of shoulder pathology, *Physical Therapy in Sport* (2014), http://dx.doi.org/10.1016/j.ptsp.2014.08.001

Table 3
Design features of the best articles reporting on diagnostic accuracy of combined tests.

Author	(Litaker et al., 2000)	(Park et al., 2005)	(Farber et al., 2006)	(Guanche & Jones, 2003)
At least 20 subjects per test in the cluster	Yes	Yes	No	Yes
Assessed conditional dependence	Yes	No	No	No
Stepwise regression	Yes	Yes	No	No
Assessed linearity of continuous measures	Yes	N/A	N/A	N/A

impingement, likely leading to diagnostic confusion between early stages of impingement and rotator cuff tears (latter stage of impingement). The 2 diagnostic clusters for labral tears, a difficult clinical diagnosis, come from a single study (Guanche & Jones, 2003) and are only moderately diagnostic when positive with positive likelihood ratios ranging between 2.67 and 5.43. Further, one of the clusters for labral tears incorporates the active compression test, a test of dubious value (Ebinger, Magosch, Lichtenberg, & Habermeyer, 2008; McFarland, Kim, & Savino, 2002; Morgan, Burkhart, Palmeri, & Gillespie, 1998; Oh, Kim, Kim, Gong, & Lee, 2008; Walsworth, Doukas, Murphy, Mielcarek, & Michener, 2008). Finally, for examining anterior instability, the results of the study (Farber et al., 2006) are likely influenced by the fact that the instability group was younger and was more likely to have a history of trauma. It is important to note that in the study by Farber et al. (2006) apprehension was used as a positive test and not reproduction of pain.

Despite the limitations of current literature on the diagnostic accuracy of test clusters to diagnose shoulder pathology, we thought it would be helpful to illustrate the best combinations through a case study (Fig. 1).

5. Case study

Our fictitious patient is 67 years old and has complained of shoulder pain of 4 months in duration. He reported that his pain initiated while walking his dogs (when they jerked the leash he held) but notes that the pain has progressed markedly over the last two months. He is able to raise his arm above his head (with pain) but has noted that his arm now aches consistently, with a more noticeable ache at night. Frequent use of ibuprofen helps modulate his pain but the effects are only temporary at best.

As a clinician, one might consider several possibilities, especially with the individual's age, consistent pain, and traumatic onset. Tests with low LR− help "rule out" competing conditions thus one might choose to consider tests or clusters of tests for shoulder labrum tears, impingement, or a rotator cuff tear. As a reminder, clustering tests often leads to higher LR+ with a sacrifice of LR−, unless the cluster was mathematically designed as a screen.

Since we have 3 competing diagnoses, attempting to rule out one or two conditions would be prudent. Currently, there are no high quality clusters of screening tests for the labrum that would rule out this condition. In the absence of a labral tear test cluster with an LR− near zero, the clinician has 2 choices: 1. consider single test results with an LR− near zero or 2. attempt to rule in the condition with a cluster of test findings that has a high LR+. Only 1 OST, the biceps load II, comes from a high quality study and has an LR− near zero (Kim, Ha, Ahn, & Choi, 2001). Unfortunately, a second high quality study (Oh et al., 2008) showed the test to have no ability to rule out (or in) a labral tear. Guanche and Jones (2003) reported a test cluster with an LR+ of 5.43 but this likelihood ratio is of only moderate assistance in diagnosing a labral tear which does not have an established set of signs and symptoms (Luime et al., 2004) and a likely low prevalence, somewhere around 6% (Snyder, Banas, & Karzel, 1995). Therefore, the best decision in this case is to treat the diagnosis of labral tear as one of exclusion and move on to rule out one of the diagnoses of either rotator cuff tear or impingement.

Park et al. (2005) reported that with impingement syndrome, negative findings of 1) Hawkins–Kennedy, 2) painful arc test and 3)

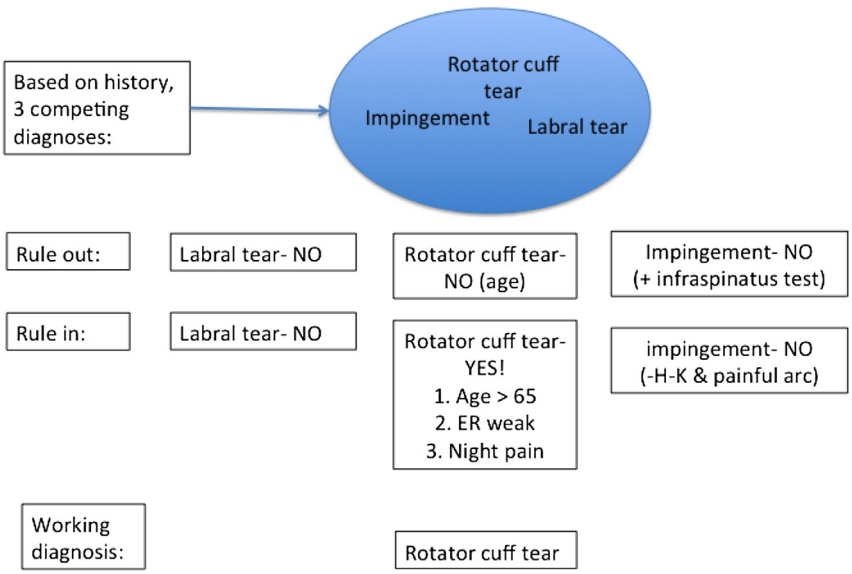

Fig. 1. Diagnostic process using the best available clinical test clusters for shoulder pathology. Key: + = positive test; − = negative test; H–K = Hawkins–Kennedy; ER = external rotation.

Please cite this article in press as: Hegedus, E. J., et al., Combining orthopedic special tests to improve diagnosis of shoulder pathology, *Physical Therapy in Sport* (2014), http://dx.doi.org/10.1016/j.ptsp.2014.08.001

infraspinatus test provides a very low LR− (0.17) and thus, has the capacity to rule out the condition. Park et al. (2005) also reported that ruling out a rotator cuff tear is possible with negative findings on 1) Age ≥ 60, 2) painful arc test, 3) drop arm test, and 4) infraspinatus test. Since our patient is over age 60, we cannot rule out a rotator cuff tear. For the sake of this case, the Hawkins–Kennedy and the painful arc tests were negative but the infraspinatus test (weakness against resisted external rotation) was positive. The impingement test cluster, therefore, cannot rule out impingement. With a negative Hawkins–Kennedy and painful arc, we also cannot rule in impingement.

To rule in a rotator cuff tear, one could refer back to the findings of Park et al. (2005) or consider the results from Litaker et al. (2000). Two consistencies with the findings of Park et al. (2005) and Litaker et al. (2000) are age and external rotation strength losses. Park et al. (2005) report the benefit of the drop arm sign and the painful arc test whereas Litaker et al. (2000) report the value of night pain. Recall that our patient is 67 years old has night pain, and a positive infraspinatus test (weakness in external rotation). These 3 findings complete the test cluster by Litaker et al. (2000). The clinician must be content with an LR+ of 9.84. Since the painful arc test was previously reported as negative, the cluster of Park et al. (2005), with an LR+ of 28.0, cannot be used.

6. Conclusion

The clinical diagnostic process should be viewed through the lens of odds and probabilities. In order to do so, test clusters from high quality studies should be utilized. In our case study, the patient likely has a rotator cuff tear but we were unable to rule out or in a labral tear and impingement. High quality clinical test clusters with powerful diagnostic characteristics for labral tears do not presently exist and impingement is an all-encompassing term for tendon pathology at the shoulder that is, at best, unhelpful in guiding treatment, and, at worst, a clinical illusion (Hegedus et al., 2012; Lewis, 2011). Other important pathologies of the shoulder like biceps tendinopathy, multi-directional instability, and fractures, also lack powerful clinical diagnostic clusters. Improved research that follows the tenets of multivariate modeling outlined in this article must be performed in order to improve the tools available to clinicians as we attempt to make use of the clinical examination to diagnose shoulder pathology.

Conflict of interest

None.

Funding

None.

Ethical approval

None.

References

Alexopoulos, E. C. (2010). Introduction to multivariate regression analysis. *Hippokratia, 14*, 23–28.

Bak, K., Sorensen, A. K., Jorgensen, U., Nygaard, M., Krarup, A. L., Thune, C., et al. (2010). The value of clinical tests in acute full-thickness tears of the supraspinatus tendon: does a subacromial lidocaine injection help in the clinical diagnosis? A prospective study. *Arthroscopy: The Journal of Arthroscopic & Related Surgery: Official Publication of the Arthroscopy Association of North America and the International Arthroscopy Association, 26*, 734–742.

Beattie, P., & Nelson, R. (2006). Clinical prediction rules: what are they and what do they tell us? *The Australian Journal of Physiotherapy, 52*, 157–163.

Beneciuk, J. M., Bishop, M. D., & George, S. Z. (2009). Clinical prediction rules for physical therapy interventions: a systematic review. *Physical Therapy, 89*, 114–124.

Chronopoulos, E., Kim, T. K., Park, H. B., Ashenbrenner, D., & McFarland, E. G. (2004). Diagnostic value of physical tests for isolated chronic acromioclavicular lesions. *The American Journal of Sports Medicine, 32*, 655–661.

Cook, C., Brown, C., Isaacs, R., Roman, M., Davis, S., & Richardson, W. (2010). Clustered clinical findings for diagnosis of cervical spine myelopathy. *The Journal of Manual & Manipulative Therapy, 18*, 175–180.

Cook, C., & Hegedus, E. J. (2013). *Orthopedic physical examination tests: An evidence-based approach* (2nd ed.). Upper Saddle River: Rearson Education, Inc.

Cook, C., Shah, A., & Pietrobon, R. (2008). Lumbopelvic manipulation for treatment of patients with patellofemoral pain syndrome: development of a clinical predication rule. *The Journal of Orthopaedic and Sports Physical Therapy, 38*, 722.

Ebinger, N., Magosch, P., Lichtenberg, S., & Habermeyer, P. (2008). A new SLAP test: the supine flexion resistance test. *Arthroscopy: The Journal of Arthroscopic & Related Surgery: Official Publication of the Arthroscopy Association of North America and the International Arthroscopy Association, 24*, 500–505.

Farber, A. J., Castillo, R., Clough, M., Bahk, M., & McFarland, E. G. (2006). Clinical assessment of three common tests for traumatic anterior shoulder instability. *The Journal of Bone and Joint Surgery, American Volume, 88*, 1467–1474.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*, 27–38.

Guanche, C. A., & Jones, D. C. (2003). Clinical testing for tears of the glenoid labrum. *Arthroscopy: The Journal of Arthroscopic & Related Surgery: Official Publication of the Arthroscopy Association of North America and the International Arthroscopy Association, 19*, 517–523.

Haskins, R., Rivett, D. A., & Osmotherly, P. G. (2012). Clinical prediction rules in the physiotherapy management of low back pain: a systematic review. *Manual Therapy, 17*, 9–21.

Hegedus, E. J., Goode, A., Campbell, S., Morin, A., Tamaddoni, M., Moorman, C. T., 3rd, et al. (2008). Physical examination tests of the shoulder: a systematic review with meta-analysis of individual tests. *British Journal of Sports Medicine, 42*, 80–92. discussion 92.

Hegedus, E. J., Goode, A. P., Cook, C. E., Michener, L., Myer, C. A., Myer, D. M., et al. (2012). 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. *British Journal of Sports Medicine, 46*, 964–978.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley and Sons.

Kim, S. H., Ha, K. I., Ahn, J. H., & Choi, H. J. (2001). Biceps load test II: a clinical test for SLAP lesions of the shoulder. *Arthroscopy: The Journal of Arthroscopic & Related Surgery: Official Publication of the Arthroscopy Association of North America and the International Arthroscopy Association, 17*, 160–164.

Kutner, M., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models* (4th ed.). Burr Ridge: McGraw-Hill/Irwin.

Lewis, J. S. (2011). Subacromial impingement syndrome: a musculoskeletal condition or a clinical illusion? *Physical Therapy Review, 16*, 388–398.

Litaker, D., Pioro, M., El Bilbeisi, H., & Brems, J. (2000). Returning to the bedside: using the history and physical examination to identify rotator cuff tears. *Journal of the American Geriatrics Society, 48*, 1633–1637.

Luime, J. J., Verhagen, A. P., Miedema, H. S., Kuiper, J. I., Burdorf, A., Verhaar, J. A., et al. (2004). Does this patient have an instability of the shoulder or a labrum lesion? *JAMA: The Journal of the American Medical Association, 292*, 1989–1999.

May, S., & Rosedale, R. (2009). Prescriptive clinical prediction rules in back pain research: a systematic review. *The Journal of Manual & Manipulative Therapy, 17*, 36–45.

McFarland, E. G., Kim, T. K., & Savino, R. M. (2002). Clinical assessment of three common tests for superior labral anterior–posterior lesions. *The American Journal of Sports Medicine, 30*, 810–815.

Menten, J., Boelaert, M., & Lesaffre, E. (2008). Bayesian latent class models with conditionally dependent diagnostic tests: a case study. *Statistics in Medicine, 27*, 4469–4488.

Michener, L. A., Walsworth, M. K., Doukas, W. C., & Murphy, K. P. (2009). Reliability and diagnostic accuracy of 5 physical examination tests and combination of tests for subacromial impingement. *Archives of Physical Medicine and Rehabilitation, 90*, 1898–1903.

Morgan, C. D., Burkhart, S. S., Palmeri, M., & Gillespie, M. (1998). Type II SLAP lesions: three subtypes and their relationships to superior instability and rotator cuff tears. *Arthroscopy: The Journal of Arthroscopic & Related Surgery: Official Publication of the Arthroscopy Association of North America and the International Arthroscopy Association, 14*, 553–565.

Nee, R. J., & Coppieters, M. W. (2011). Interpreting research on clinical prediction rules for physiotherapy treatments. *Manual Therapy, 16*, 105–108.

Oh, J. H., Kim, J. Y., Kim, W. S., Gong, H. S., & Lee, J. H. (2008). The evaluation of various physical examinations for the diagnosis of type II superior labrum anterior and posterior lesion. *The American Journal of Sports Medicine, 36*, 353–359.

Park, H. B., Yokota, A., Gill, H. S., El Rassi, G., & McFarland, E. G. (2005). Diagnostic accuracy of clinical tests for the different degrees of subacromial impingement syndrome. *The Journal of Bone and Joint Surgery, American Volume, 87*, 1446–1455.

Shen, J., & Gao, S. A. (2008). Solution to separation and multicollinearity in multiple logistic regression. *Journal of Data Science, 6*, 515–531.

Snyder, S. J., Banas, M. P., & Karzel, R. P. (1995). An analysis of 140 injuries to the superior glenoid labrum. *Journal of Shoulder and Elbow Surgery/American Shoulder and Elbow Surgeons… [et al.], 4*, 243–248.

Stanton, T. R., Hancock, M. J., Maher, C. G., & Koes, B. W. (2010). Critical appraisal of clinical prediction rules that aim to optimize treatment selection for musculoskeletal conditions. *Physical Therapy, 90*, 843–854.

Please cite this article in press as: Hegedus, E. J., et al., Combining orthopedic special tests to improve diagnosis of shoulder pathology, *Physical Therapy in Sport* (2014), http://dx.doi.org/10.1016/j.ptsp.2014.08.001

Walsworth, M. K., Doukas, W. C., Murphy, K. P., Mielcarek, B. J., & Michener, L. A. (2008). Reliability and diagnostic accuracy of history and physical examination for diagnosing glenoid labral tears. *The American Journal of Sports Medicine, 36*, 162–168.

Walton, J., Mahajan, S., Paxinos, A., Marshall, J., Bryant, C., Shnier, R., et al. (2004). Diagnostic values of tests for acromioclavicular joint pain. *The Journal of Bone and Joint Surgery, American Volume, 86-A*, 807–812.

Whiting, P., Rutjes, A. W., Dinnes, J., Reitsma, J., Bossuyt, P. M., & Kleijnen, J. (2004). Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technology Assessment, 8*(iii), 1–234.

Wlikinson, L., & Dallal, G. E. (1981). Tests of significance in forward selection regression with an F-to enter stopping rule. *Technometrics, 23*, 377–380.

Please cite this article in press as: Hegedus, E. J., et al., Combining orthopedic special tests to improve diagnosis of shoulder pathology, *Physical Therapy in Sport* (2014), http://dx.doi.org/10.1016/j.ptsp.2014.08.001