Creation and Critique of Studies of Diagnostic Accuracy: Use of the STARD and QUADAS Methodological Quality Assessment Tools

Chad Cook, PT, PhD, MBA, OCS, FAAOMPT
Joshua Cleland, PT, DPT, PhD, OCS, FAAOMPT
Peter Huijbregts, PT, DPT, OCS, FAAOMPT, FCAMT

Abstract: Clinical special tests are a mainstay of orthopaedic diagnosis. Within the context of the evidence-based practice paradigm, data on the diagnostic accuracy of these special tests are frequently used in the decision-making process when determining the diagnosis, prognosis, and selection of appropriate intervention strategies. However, the reported diagnostic utility of these tests is significantly affected by study methodology of diagnostic accuracy studies. Methodological shortcomings can influence the outcome of such studies, and this in turn will affect the clinician's interpretation of diagnostic findings. The methodological issues associated with studies investigating the diagnostic utility of clinical tests have mandated the development of the STARD (Standards for Reporting of Diagnostic Accuracy) and QUADAS (Quality Assessment of Diagnostic Accuracy Studies) criterion lists. The purpose of this paper is to outline the STARD and QUADAS criterion lists and to discuss how these methodological quality assessment tools can assist the clinician in ascertaining clinically useful information from a diagnostic accuracy study.

Key Words: Special Tests, Diagnostic Accuracy, Methodological Quality Assessment Tools, STARD, QUADAS

The clinician's armamentarium for screening, diagnosis, and prognosis of selected conditions has expanded with the creation of numerous clinical special tests. Diagnostic tests are decidedly dynamic, as new tests are developed concurrently with improvements in technology[1]. These clinical tests remain extremely popular among orthopaedic diagnosticians, and information gained from these tests is frequently considered during decision-making with regard to patient diagnosis and prognosis and the selection of appropriate interventions. Historically, textbooks describing these tests have ignored mention of the tests' true ability to identify the presence of a particular disorder as based on studies into diagnostic utility of these tests; rather, they have concentrated solely on test description and scoring. Relying solely on a pathophysiologic and/or pathobiomechanical rationale for the interpretation of clinical tests without considering the research data on diagnostic accuracy of said tests can potentially result in the selection of tests that provide little worthwhile diagnostic or prognostic information. In addition, it can lead clinicians to make incorrect treatment decisions[1]. With the number of clinical special tests and measures continuing to multiply, it is essential to thoroughly evaluate a test's diagnostic utility prior to incorporating it into clinical practice[2,3].

Clinical special tests exhibit the measurable diagnostic properties of sensitivity and specificity. The sensitivity of a test is the ability of the test to identify a positive finding when the targeted diagnosis is actually present (i.e., true positive)[4]. Specificity is the discriminatory ability of a test to identify if the disease or condition is absent when in actuality it is truly is absent (i.e., true negative)[4]. Sensitivity and specificity values can then be used to calculate positive and negative likelihood ratios (LR). Although sensitivity and specificity—when high—are useful for confirming the presence or absence of a specific disorder, the general consensus seems to be that likelihood

Address all correspondence and requests for reprints to:
Chad Cook
Assistant Professor
Director of Outcomes Research
Center for Excellence in Surgical Outcomes
Department of Surgery
Duke University
Durham, NC 27710
E-mail: chad.cook@duke.edu

ratios are the optimal statistics for determining a shift in the pretest probability that a patient has a specific disorder. Table 1 provides information on statistics relevant to diagnostic utility.

Clinical special tests that demonstrate strong sensitivity are considered clinically useful screening tools[5] in that they can be used for ruling out selected diagnoses or impairments[6]. When a test demonstrates high sensitivity, the likelihood of a false negative finding (i.e., incorrectly identifying the patient as not having the disorder when in reality she actually does have said condition) is low since the test demonstrates the substantive ability to identify those who truly have the disease or impairment, thus demonstrating the ability to "rule out" a condition. Conversely, tests that demonstrate high specificity are appropriate for "ruling in" a finding, indicating that a positive value is more telling than a negative value. The likelihood of a false positive is low because the test demonstrates the ability to correctly identify those who truly do not have the disease or impairment. This ability of highly sensitive and highly specific tests to rule in a condition or rule out a condition, respectively, is captured in the mnemonic below:

- **SnNOUT**: With highly **Sn**sitive tests, a **N**egative result will rule a disorder **OUT**
- **SpPIN**: With highly **Sp**ecific tests, a **P**ositive result will rule a disorder **IN**

Likelihood ratios can be either positive or negative. A positive likelihood ratio (LR+) indicates a shift in probability favoring the existence of a disorder if the test is found to be positive. A value of 1 indicates an equivocal strength of diagnostic power; values that are higher suggest greater strength. Conversely, a negative likelihood ratio (LR–) indicates a shift in probability favoring the absence of a disorder if the test is found to be negative. The lower the value, the better the ability of the test to determine the post-test odds that the disease is actually absent in the event the finding is negative. A number closer to 1 indicates that a negative test is equally likely to occur in individuals with or without the disease. Table 2 represents the shifts in probability associated with specific range of positive and negative likelihood ratios that a patient does or does not have a particular disorder given a positive or negative test[7].

With the intent of providing a comprehensive overview of all statistical measures relevant to diagnostic utility, Table 1 also provides definitions for three additional statistics. The accuracy of a diagnostic test provides a quantitative measure of its overall value, but because it does not differentiate between the diagnostic value of positive and negative test results, its value with regard to diagnostic decisions is minimal. At first sight, positive and negative predictive values seem to have greater diagnostic value. However, because the prevalence in the clinical population being examined has to be identical to the prevalence in the study population from which the predictive values were derived before we can justifiably use predictive values as a basis for diagnostic decisions, their usefulness is again limited.

Many orthopaedic clinical tests are products of traditional examination methods and principles; i.e., the tests

TABLE 1. Definition and calculation of statistical measures used to express diagnostic test utility

Statistical measure	Definition	Calculation
Accuracy	The proportion of people who were correctly identified as either having or not having the disease or dysfunction	(TP + TN) / (TP + FP + FN + TN)
Sensitivity	The proportion of people who have the disease or dysfunction who test positive	TP / (TP + FN)
Specificity	The proportion of people who do not have the disease or dysfunction who test negative	TN / (FP + TN)
Positive predictive value	The proportion of people who test positive and who have the disease or dysfunction	TP / (TP + FP)
Negative predictive value	The proportion of people who test negative and who do not have the disease or dysfunction	TN / (FN + TN)
Positive likelihood ratio	How likely a positive test result is in people who have the disease or dysfunction as compared to how likely it is in those who do not have the disease or dysfunction)	Sensitivity/(1-specificity)
Negative likelihood ratio	How likely a negative test result is in people who have the disease or dysfunction as compared to how likely it is in those who do not have the disease or dysfunction	(1-sensitivity)/specificity

TP= true positive; TN= true negative; FP= false positive; FN= false negative

TABLE 2. Diagnostic value guidelines

LR+	Interpretation
> 10	Large and often conclusive increase in the likelihood of disease
5 - 10	Moderate increase in the likelihood of disease
2 - 5	Small increase in the likelihood of disease
1 - 2	Minimal increase in the likelihood of disease
1	No change in the likelihood of disease
LR–	**Interpretation**
1	No change in the likelihood of disease
0.5 - 1.0	Minimal decrease in the likelihood of disease
0.2 - 0.5	Small decrease in the likelihood of disease
0.1 - 0.2	Moderate decrease in the likelihood of disease
< 0.1	Large and often conclusive decrease in the likelihood of disease

LR+ = Positive Likelihood Ratio
LR– = Negative Likelihood Ratio

were based solely on a pathophysiologic and/or pathobiomechanical rationale. For example, Spurling and Scoville introduced the Spurling's sign in 1944 as a diagnostic test for cervical radiculopathy[8]. Over 125 different articles advocating the merit of this test as a diagnostic tool have since cited this 1944 study. In the original article, Spurling and Scoville[8] reported a sensitivity of 100%, exclusively for identifying patients with cervical radiculopathy. Consequently, Spurling's maneuver has been frequently used as a tool for screening for or, in some cases, diagnosing cervical radiculopathy and cervical herniated disks[9-13]. Table 3 outlines the findings of a number of studies that have investigated the diagnostic utility of the Spurling's test. The findings in this table can be used to illustrate an important point: Despite the claims by Spurling and Scoville of perfect sensitivity for the Spurling's test with regard to identifying the presence of cervical radiculopathy, subsequent studies that have investigated the diagnostic value of Spurling's maneuver have found dramatically different results from those initially reported. For example, Uchihara et al[9] reported that the Spurling test exhibited a sensitivity of 11% and a specificity of 100%, while Tong et al[14] reported a sensitivity of 30% and a specificity of 93%. Additional researchers[11-12] have found sensitivities similar to those reported in these two studies[9,14]. However, none have reported values near 100%. Since the numbers among studies are dramatically different, clinicians are left with the following question: Is the test more appropriately used as a screening tool as advocated by Spurling and Scoville[8] or as a measure of fair to moderate diagnostic utility as suggested by a number of other authors?

The answer lies in the methodological rigor in the study design and the applicability of these findings to the diagnostic environment of the practicing clinician. Methodological issues can influence the outcome of diagnostic utility studies, and this in turn should affect the clinician's interpretation of diagnostic findings. The methodological issues associated with studies investigating the diagnostic utility of clinical tests have mandated the development of criterion lists to systematically determine the methodological quality of diagnostic utility studies, i.e., the STARD (Standards for Reporting of Diagnostic Accuracy) and QUADAS (Quality Assessment of Diagnostic Accuracy Studies) criteria. The purpose of this paper is to outline the STARD and QUADAS criteria and to discuss how these criteria can assist the clinician in ascertaining clinically useful information from a diagnostic accuracy study.

Common Design Errors in Diagnostic Accuracy Studies

A rigorous evaluation process of the true accuracy of clinical special tests can reduce the rate of erroneous diagnostic results[15]. Exaggerated reports of diagnostic utility from poorly designed studies can potentially result in an inaccurate diag-

TABLE 3. Diagnostic accuracy studies of the Spurling's maneuver

Study	Sensitivity (%)	Specificity (%)	+LR	–LR
Spurling & Scoville[8]	100	NT	NA	NA
Uchihara et al[9]	11	100	NA	NA
Shah & Rajshekhar[10]	93.1	95	18.6	0.07
Wainner et al[11]	50	86	3.57	0.58
Wainner et al[11] (Test included side flexion towards the rotation and extension)	50	74	1.92	0.67
Viikari-Juntura et al[12] (Right side)	36	92	4.5	0.69
Viikari-Juntura et al[12] (Left side)	39	92	4.87	0.66
Sandmark & Nisell[13] (Not for radiculopathy)	77	92	9.62	0.25
Tong et al[14]	30	93	4.3	0.75

nosis, inappropriate treatment, premature adoption of a special test that provides little value, and errors in clinical decision-making[16]. Past studies have suggested that the methodological qualities of many studies designed to investigate the diagnostic utility of clinical tests are mediocre at best[16]. Numerous methodological shortcomings are apparent in the design of diagnostic accuracy studies. Common methodological shortcomings relate to:

- Use of an inappropriate gold standard or reference test
- Spectrum or selection bias
- Lack of rater blinding
- Insufficient operational definition of positive and negative test findings
- Absence of a third category of indeterminate test findings

When investigating the accuracy of a special test, the test under investigation is compared to a gold standard or reference test (criterion test) that is considered the best available representation of the truth with regard to the condition of interest (i.e., the reference test is expected to identify all those with the disorder)[17,18]. In most instances, the optimal reference test consists of findings during surgical intervention. In cases where the reference standard is disputed, controversial, or difficult to identify, authors have suggested the use of a clinical diagnosis of similar signs and symptoms, often predicated on the erroneous assumption that the clinical tests selected as a reference standard are explicit to the disorder that is being measured. For example, Cibulka and Koldehoff[19] have suggested the use of a cluster of purported sacroiliac tests to define those who do and do not exhibit symptoms of low back pain. This suggests that the cluster of tests offers clinical utility without the need for the currently most used reference standard for the identification of sacroiliac pathology, i.e., a sacroiliac joint injection. Studies such as these can thereby artificially inflate the identified diagnostic accuracy findings, and studies using this form of a reference standard run the risk of significant bias that can result in inaccurate findings.

Second, many diagnostic utility studies suffer from spectrum or selection bias[4]. Spectrum or selection bias occurs when the subjects of a particular study are not representative of the population to which the test is generally applied. Consider the situation in which the population that is tested in the study consists of those subjects who have a high prevalence of a specific condition inferring a high likelihood that the disorder is present. This may occur when studies are performed in specialized secondary centers, such as specific pain centers or surgical offices. This causes the likelihood that the test will be found positive to exceed that of the population that might be suspected of having the condition studied and that would typically present to primary centers such as physical therapy or chiropractic clinics and primary care physician's offices. Often, spectrum bias will overtly improve the sensitivity of a test and inaccurately inflate the diagnostic value. The biceps load test created by Kim et al[20] provides an example of spectrum bias that inflated the sensitivity of the test and thereby artificially heightened the LR+. Kim et al[20] evaluated the ability of the biceps load test to identify the presence of a labral tear in 75 successive patients with unilateral recurrent anterior shoulder dislocations and a Bankart lesion. The presence of a labral tear in a population of patients with a history of anterior dislocation would be expected to be significantly greater than in a general population presenting with reports of shoulder pain; thus, spectrum bias clearly influenced the findings of this study.

Third, numerous studies that have investigated the diagnostic utility of special tests lack appropriate rater blinding. Consequently, clinicians may have a predisposition to select a positive or negative consequence based on knowledge of the results of the reference test findings, other additional diagnostic information, past experience, or personal preference. Glaser et al[21] reported the diagnostic accuracy findings of the Hoffmann's test in diagnosing patients with suspected myelopathy. When blinded to other components of the examination, the authors reported a sensitivity of 28% and a specificity of 71% (LR+ = 0.96; LR– = 1.01). When the raters were not blinded, the findings increased to a sensitivity of 58% and a specificity of 74% (LR+= 2.2; LR– = 0.57). Although one may argue that myelopathy requires the understanding of multiple findings—and thus the necessity of not blinding raters—to make a correct diagnosis, this study suggests that the likelihood of reporting a positive finding for the Hoffmann's sign was significantly affected by other factors, such as patient history, which lie outside the findings of the individual test. Thus, the diagnostic accuracy of the specific test results under investigation may again be artificially inflated.

Fourth, many tests lack an appropriate threshold or cut-off score to signify either a positive or negative finding. Altering the cut-off point that determines whether a test is positive or negative can significantly affect the sensitivity and specificity of a test; for the study results to be applied to the clinical situation, operational definitions of positive and negative test findings on dichotomous tests need to be similar.

Finally, in some instances, the results of a special clinical test are inconclusive and do not yield findings that are above or below the threshold and thereby provide only limited clinical usefulness. For example, the straight leg raise (SLR) test is commonly used to determine the presence of lumbar radiculopathy or neural tension[23-25]. However, the operational definition describing whether a test is positive or negative is often disputed[26-27], and in some instances what appears as a positive SLR test may not actually signify the presence of lumbar radiculopathy[28-29]. Subsequently, a finding of *indeterminate* is an appropriate selection in the many

cases where it is difficult to identify if a special test is truly either positive or negative. Because the failure to find a positive or negative test can significantly affect the diagnostic usefulness of a test, this consequence should be identified in special clinical tests studies. Therefore, in the development of one of the methodological quality assessment tools discussed below, Bossuyt et al[22] recommended the use of a third category to define tests that are indeterminate, which would provide for a more accurate representation of a test's true diagnostic accuracy.

These examples identify some potential biases associated with establishing the diagnostic utility of clinical special tests. The numerous methodological issues associated with studies investigating the diagnostic utility of clinical tests clearly mandate that specific guidelines are necessary to assist with carrying out and critiquing such a study. These methodological issues have resulted in the development of two separate methodological quality assessment tools: the STARD (Standards for Reporting of Diagnostic Accuracy) and the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) criterion lists.

Methodological Quality Assessment Tools

STARD Criterion List

In 1999, the Cochrane Diagnostic and Screening Test Methods Working Group met at the Cochrane Colloquium in Rome, Italy[30]. Following the premise set by the CONSORT (Consolidated Standards of Reporting Trials) initiative to develop methodological standards for a specific study design, the

Fig. 1. Flow Chart for the STARD (Standards for Reporting Diagnostic Accuracy) checklist.

group developed a checklist of items by way of expert consensus with the focus on improving the design of studies investigating the diagnostic accuracy of tests or measures. The workgroup developed the Standards for Reporting of Diagnostic Accuracy (STARD) checklist, a 25-item checklist created by narrowing the results from an extensive search of the literature that revealed the presence of 33 methodological scoring lists for studies on diagnostic accuracy with 75 distinct individual items[31,32].

Similar to the CONSORT standards, the STARD checklist is designed to provide researchers with a checklist and flow diagram (Figure 1) that should be followed for optimal study design. The flow diagram outlines a method for patient recruitment, the order of test execution, the number of patients undergoing the index test, and the specific reference test selected. Each flow phase includes a section for an inconclusive test finding (neither positive nor negative) in the index or reference test and provides a venue by which this can be identified.

The STARD checklist is divided into five major topics and six subsequent subtopics. The five major topics—1) title, abstract, and keywords; 2) introduction; 3) methods; 4) results; and 5) discussion—provide suggestions for study design and reporting of the results to improve the reader's ability to identify and judge the methodological rigor of diagnostic accuracy studies. Each of the six subsequent subtopics further break down the elemental design for participants, test methods and application, statistical methods, and estimates of diagnostic accuracy. By using the checklist during the design phase, researchers are less apt to incorporate fatal study errors such as recruitment or selection bias or overlap between the index test and reference test findings, and readers are more likely to improve their critique of diagnostic accuracy study secondary to standardization of reporting and design. Table 4 outlines the STARD checklist.

QUADAS Criterion List

Recently, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool was developed to assess the quality (both internal and external) of primary research studies of diagnostic accuracy[33]. QUADAS was developed through a four-round Delphi panel, which reduced 28 critical criteria for evaluation of a completed diagnostic accuracy study to 14 final components. Within the 14 components outlined in Table 5, three overall criteria that are assessed include 1) reporting of selection criteria, 2) description of index test execution, and 3) description of reference standard execution. Each of the 14 steps is scored as "yes," "no," or "unclear." Whiting et al[3] provided individual procedures for scoring each of the 14 items, including operational standards for each question. In research terms, the QUADAS tool provides an organized format in which readers can examine the internal validity and external validity of a study.

Current Research Status of STARD and QUADAS Tools

Nonetheless, at present, neither QUADAS nor STARD is used in quantifying a value or score for diagnostic accuracy[34]. At best, systematic reviews that use the QUADAS instrument provide a qualitative assessment of design with recognition that weaknesses in selected regions may alter some test findings more than others. However, recent interrater reliability testing of the QUADAS has demonstrated adequate agreement for individual items in the checklist (range 50–100%, median 90%)[3].

Intent of the STARD and QUADAS Tools

The QUADAS and STARD differ from each other in the intent of the instrument. While STARD is a prospective tool used to outline the development of a well-designed study, QUADAS is considered a retrospective instrument used to critique the methodological rigor of a study investigating the diagnostic accuracy of a test or measure. QUADAS is designed to serve as a critique of a completed study and a measure of the study design, and it incorporates many of the same items provided in the checklist by the STARD initiative. The principal purpose of the STARD is to improve the quality of reporting of diagnostic studies. STARD is designed to outline the specific features required for an unbiased diagnostic accuracy study and to improve the ability of the reader to gauge the strength of a finding through commonality in reporting.

Clinical Application

Despite the difference in intent of the tools, both STARD and QUADAS can be useful for clinicians when determining the strength of a study and hence the value of a selected clinical test. For example, because tests of high sensitivity are useful in ruling out the presence of a condition, incorporating these tests early in the examination is helpful in organizing a thoughtful progression of the remainder of the examination. However, artificially low or high sensitivity scores reported by studies with less than optimal scientific rigor may bias the clinician's findings potentially resulting in an inaccurate diagnosis. The various reported accuracies of the Spurling's test are examples of potentially misleading findings in that some studies have found the test to be highly sensitive but not specific while others have found exactly the opposite. When reviewed using the QUADAS criteria, the studies[8,10,13] that found Spurling's test to be sensitive had specific components of the QUADAS criteria missing or findings that could increase the risk of selection bias, while those that found the test specific[11,12] exhibited fewer missing criteria and greater integrity in selection and interpretation standards (Table 6). This suggests that the studies that found higher levels of

TABLE 4. The STARD (Standards for Reporting Diagnostic Accuracy) checklist

Manuscript number and/or corresponding author name:

TITLE /ABSTRACT/KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.
METHODS		Describe
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where the data were collected.
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?
Test methods	7	The reference standard and its rationale
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.
	9	Definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard.
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other tests and describe any other clinical information available to the readers.
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g., 95% confidence intervals).
	13	Methods for calculating test reproducibility, if done.
RESULTS		Report
Participants	14	When study was done, including beginning and ending dates of recruitment.
	15	Clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).
	16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).
Test Results	17	Time interval from the index tests to the reference standard, and any treatment administered between.
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.
	19	A cross-tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.
	20	Any adverse events from performing the index tests or the reference standard.
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals).
	22	How indeterminate results, missing responses and outliers of the index tests were handled.
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.
	24	Estimates of test reproducibility, typically imprecision (as CV) at 2 or 3 concentrations.
DISCUSSION	25	Discuss the clinical applicability of the study findings.

TABLE 5. The QUADAS (Quality Assessment of Diagnostic Accuracy Studies) assessment tool

Item	Yes	No	Unclear
1 Was the spectrum of patients representative of the patients who will receive the test in practice?			
2. Were selection criteria clearly described?			
3. Is the reference standard likely to classify the target condition correctly?			
4. Is the period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?			
5. Did the whole sample or random selection of the sample receive verification using a reference standard of diagnosis?			
6. Did patients receive the same reference standard regardless of the index test result?			
7. Was the reference standard independent of the index test (i.e., the index test did not form part of the reference standard)?			
8. Was the execution of the index test described in sufficient detail to permit its replication?			
9. Was the execution of the reference standard described in sufficient detail to permit its replication?			
10. Were the index test results interpreted without knowledge of the results of the reference test?			
11. Were the reference standard results interpreted without knowledge of the results of the index test?			
12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?			
13. Were uninterpretable/intermediate test results reported?			
14. Were withdrawals from the study explained?			

specificity were of greater quality and exhibited findings more applicable to clinical practice. Subsequently, it appears that the Spurling's test is not a sensitive test and should not be used as a screening tool. Based on a systematic assessment of the methodological quality of the studies into the diagnostic utility of the Spurling's test, other tests need to be considered to more effectively "rule out" cervical radiculopathy.

Of course, within clinical practice, single tests are seldom used as the sole determinant for establishing a diagnosis. Rather, recent orthopaedic research has emphasized the use of test clusters to establish a diagnosis[11] and—with the increasing emphasis on the development of clinical prediction rules within orthopaedic manual physical therapy[34-38]—an appropriate course of management. These diagnostic test clusters and clinical prediction rule studies can and should, of course, also be assessed for methodological quality using these same assessment tools (STARD and QUADAS) with similar implications for their use in clinical practice.

Conclusion

The STARD and the QUADAS tools were developed to improve the construction and reporting (STARD) and the assessment (QUADAS) of diagnostic accuracy studies. Improvement in the study design and reporting of diagnostic accuracy studies should positively influence the quality of data available to the clinician on the diagnostic utility of tests used for screening, diagnosis, prognosis, and treatment planning. Clinicians can improve their research-based confidence in their interpretation of findings on clinical tests and measures by recognizing common biases in diagnostic accuracy studies as discussed above and by familiarizing themselves with and applying the STARD and QUADAS tools to said studies. ■

TABLE 6. QUADAS (Quality Assessment of Diagnostic Accuracy Studies) values of the Spurling test for multiple studies.

Item	Spurling & Scoville[8]	Uchihara et al[9]	Shah & Rajshekhar[10]	Wainner et al[11]	Viikari-Juntura et al[12]	Sandmark & Nisell[13]
1. Was the spectrum of patients representative of the patients who will receive the test in practice?	Y	Y	Y	Y	Y	N
2. Were selection criteria clearly described?	N	N	Y	Y	Y	Y
3. Is the reference standard likely to classify the target condition correctly?	Y	Y	Y	Y	Y	N
4. Is the period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	U	U	N	U	Y	Y
5. Did the whole sample or random selection of the sample receive verification using a reference standard of diagnosis?	U	Y	Y	Y	Y	Y
6. Did patients receive the same reference standard regardless of the index test result?	U	Y	Y	Y	Y	Y
7. Was the reference standard independent of the index test (i.e., the index test did not form part of the reference standard)?	U	Y	Y	Y	Y	N
8. Was the execution of the index test described in sufficient detail to permit its replication?	Y	Y	Y	Y	Y	Y
9. Was the execution of the reference standard described in sufficient detail to permit its replication?	N	N	N	Y	N	Y
10. Were the index test results interpreted without knowledge of the results of the reference test?	U	U	U	Y	Y	Y
11. Were the reference standard results interpreted without knowledge of the results of the index test?	U	Y	Y	U	Y	Y
12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	Y	Y	Y	Y	Y	Y
13. Were uninterpretable/intermediate test results reported?	N	N	N	U	N	N
14. Were withdrawals from the study explained?	N	N	N	N	N	N

Y = Yes, N = No, U = Unclear

REFERENCES

1. Bossuyt PMM. The quality of reporting in diagnostic test research: Getting better, still not optimal. *Clin Chem* 2004;50:465–467.
2. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282: 1061–1063.
3. Whiting PF, Weswood ME, Rutjes AW, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* 2006;6:9.
4. Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: A single indicator of test performance. *J Clin Epidemiol* 2003;56:1129–1135.
5. Woolf SF. Do clinical practice guidelines define good medical care? The need for good science and the disclosure of uncertainty when defining "best practices." *Chest* 1998;113(3 Suppl):166S–171S.
6. Sackett DL, Straus S, Richardson S, et al. *Evidence-Based Medicine: How to Practice and Teach EBM*. 2nd ed. London, England: Churchill Livingstone, 2000.
7. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA* 1994;271:389–391.
8. Spurling RG, Scoville WB. Lateral rupture of the cervical intervertebral disc. *Surg Gynecol Obstet* 1944;78:350–358.
9. Uchihara T, Furukawa T, Tsukagoshi H. Compression of brachial plexus as a diagnostic test of cervical cord lesion. *Spine* 1994;19: 2170–2173.
10. Shah KC, Rajshekhar V. Reliability of diagnosis of soft cervical disc prolapse using Spurling's test. *Br J Neurosurg* 2004;18:480–483.
11. Wainner RS, Fritz JM, Irrgang JJ, et al. Reliability and diagnostic accuracy of the clinical examination and patient self-report measures for cervical radiculopathy. *Spine* 2003;28:52–62.
12. Viikari-Juntura E, Porras M, Lassonen EM. Validity of clinical tests in the diagnosis of root compression in cervical disc disease. *Spine* 1989;14:253–257.
13. Sandmark H, Nisell R. Validity of five common manual neck pain provoking tests.*Scand J Rehabil Med* 1995;27:131–136.
14. Tong HC, Haig AJ, Yamakawa K. The Spurling test and cervical radiculopathy. *Spine* 2002;27:156–159.
15. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. The Standards for Reporting of Diagnostic Accuracy Group. *Croat Med J* 2003;44:639–650.
16. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. The Standards for Reporting of Diagnostic Accuracy Group. *Croat Med J* 2003;44:635–638.
17. Jaeschke R, Meade M, Guyatt G, Keenan SP, Cook DJ. How to use diagnostic test articles in the intensive care unit: Diagnosing weanability using f/vt. *Crit Care Med* 1997;25:1514–1521.
18. Fritz JM, Wainner RS. Examining diagnostic tests: An evidence-based perspective. *Phys Ther* 2001;81(9):1546–1564.
19. Cibulka MT, Kodekoff R. Clinical usefulness of a cluster of sacroiliac joint tests in patients with and without low back pain. *J Orthop Sports Phys Ther* 1999;29:83–89.
20. Kim SH, Ha KI, Han KY. Biceps load test: A clinical test for superior labrum anterior and posterior lesions in shoulders with recurrent anterior dislocations. *Am J Sports Med* 1999;27:300–303.
21. Glaser J, Cure J, Bailey K, Morrow D. Cervical spinal cord compression and the Hoffmann sign. *Iowa Orthop J* 2001;21:49–52.
22. Bossuyt P, Reitsma J, Bruns D, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Family Practice* 2004;21:4–10.
23. Lew PC, Briggs CA. Relationship between the cervical component of the slump test and change in hamstring muscle tension. *Man Ther* 1997; 2:98–105.
24. Johnson EK, Chiarello CM. The slump test: The effects of head and lower extremity position on knee extension. *J Orthop Sports Phys Ther* 1997; 26:310–317.
25. Lew PC, Morrow CJ, Lew AM. The effect of neck and leg flexion and their sequence on the lumbar spinal cord: Implications in low back pain and sciatica. *Spine* 1994; 19:2421–2424.
26. Cameron DM, Bohannon RW, Owen SV. Influence of hip position on measurements of the straight leg raise test. *J Orthop Sports Phys Ther* 1994;19:168–172.
27. Hall TM, Elvey RL. Nerve trunk pain: Physical diagnosis and treatment. *Man Ther* 1999;4:63–73.
28. Charnley, J. Orthopaedic signs in the diagnosis of disc protrusion with special reference to the straight-leg-raising test. *Lancet* 1951;1:186–192.
29. McCombe PF, Fairbank JT, Cockersole BC, Pynsent PB. Reproducibility of physical signs in low-back pain. *Spine* 1989;14:908–918.
30. Bossyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Ann Intern Med* 2003;138:W1–W12.
31. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Clin Chem* 2003;49:1–6.
32. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clin Chem* 2003;49:7–18.
33. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol* 2005;5:19.
34. Flynn T, Fritz J, Whitman J, Wainner R, Magel J, Rendeiro D, et al. A clinical prediction rule for classifying patients with low back pain who demonstrate short-term improvement with spinal manipulation. *Spine* 2002;27:2835–2843.
35. Childs JD, Fritz JM, Flynn TW, Irrgang JJ, Johnson KK, Majkowski GR, Delitto A. A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: A validation study. *Ann Intern Med* 2004;141:920–928.
36. Fritz JM, Childs JD, Flynn TW. Pragmatic application of a clinical prediction rule in primary care to identify patients with low back pain with a good prognosis following a brief spinal manipulation intervention. *BMC Family Practice* 2005;6:29.
37. Tseng YL, Wang WTJ, Chen WY, Hou TJ, Chen TC, Lieu FK. Predictors for the immediate responders to cervical manipulation in patients with neck pain. *Man Ther* 2006;11:306–315.
38. Cleland JA, Childs JD, Fritz JM, Whitman JM, Eberhart SL. Development of a clinical prediction rule for guiding treatment of a subgroup of patients with neck pain: Use of thoracic spine manipulation, exercise, and patient education. *Phys Ther* 2007;87:9–23.