Examining Diagnostic Tests: An Evidence-Based Perspective

Diagnosis is an important aspect of physical therapist practice. Selecting tests that will provide the most accurate information and evaluating the results appropriately are important clinical skills. Most of the discussion in physical therapy to date has centered on defining diagnosis, with considerably less attention paid to elucidating the diagnostic process. Determining the best diagnostic tests for use in clinical situations requires an ability to appraise evidence in the literature that describes the accuracy and interpretation of the results of testing. Important issues for judging studies of diagnostic tests are not widely disseminated or adhered to in the literature. Lack of awareness of these issues may lead to misinterpretation of the results. The application of evidence to clinical practice also requires an understanding of evidence and its use in decision making. The purpose of this article is to present an evidence-based perspective on the diagnostic process in physical therapy. Issues relevant to the appraisal of evidence regarding diagnostic tests and integration of the evidence into patient management are presented. [Fritz JM, Wainner RS. Examining diagnostic tests: an evidence-based perspective. *Phys Ther.* 2001;81:1546–1564.]

Key Words: *Diagnosis, Evidence-based practice, Likelihood ratio, Sensitivity, Specificity.*

Julie M Fritz, Robert S Wainner

Physical therapy has a rich history of dialogue concerning the meaning of diagnosis within the profession.[1,2] The *Guide to Physical Therapist Practice* (2nd ed)[3] (the Guide) identifies diagnosis as 1 of 5 interrelated elements of patient management (Fig. 1). The Guide describes diagnosis as composed of 2 aspects: first, the process of evaluating data obtained from the examination, and, second, the end result of such a process.[3] As noted by Delitto and Snyder-Mackler[4] in 1995, debate has focused mostly on clarifying the role and function of the end result of the diagnostic process, with little attention devoted to the first aspect of the Guide's definition: the process of diagnosis. Since the time this observation was made, the paucity of discourse on the diagnostic process has persisted. The Guide identifies diagnosis as a keystone in the process of maximizing patient outcomes, representing the culmination of the examination and evaluation process and directing subsequent decisions related to prognosis and interventions. In view of the role that diagnosis is given in the Guide and the identification of numerous priorities related to diagnosis within the Clinical Research Agenda for Physical Therapy of the American Physical Therapy Association (APTA),[5] we believe that the need for further discussion of the diagnostic process is of paramount professional importance. The purpose of this article is to describe the diagnostic process in physical therapy from an evidence-based perspective. Issues relevant to the appraisal of evidence regarding diagnostic tests are presented, and the integration of evidence into clinical practice is discussed.

Diagnosis serves as the link between examination findings and interventions.

The Diagnostic Process in Physical Therapy

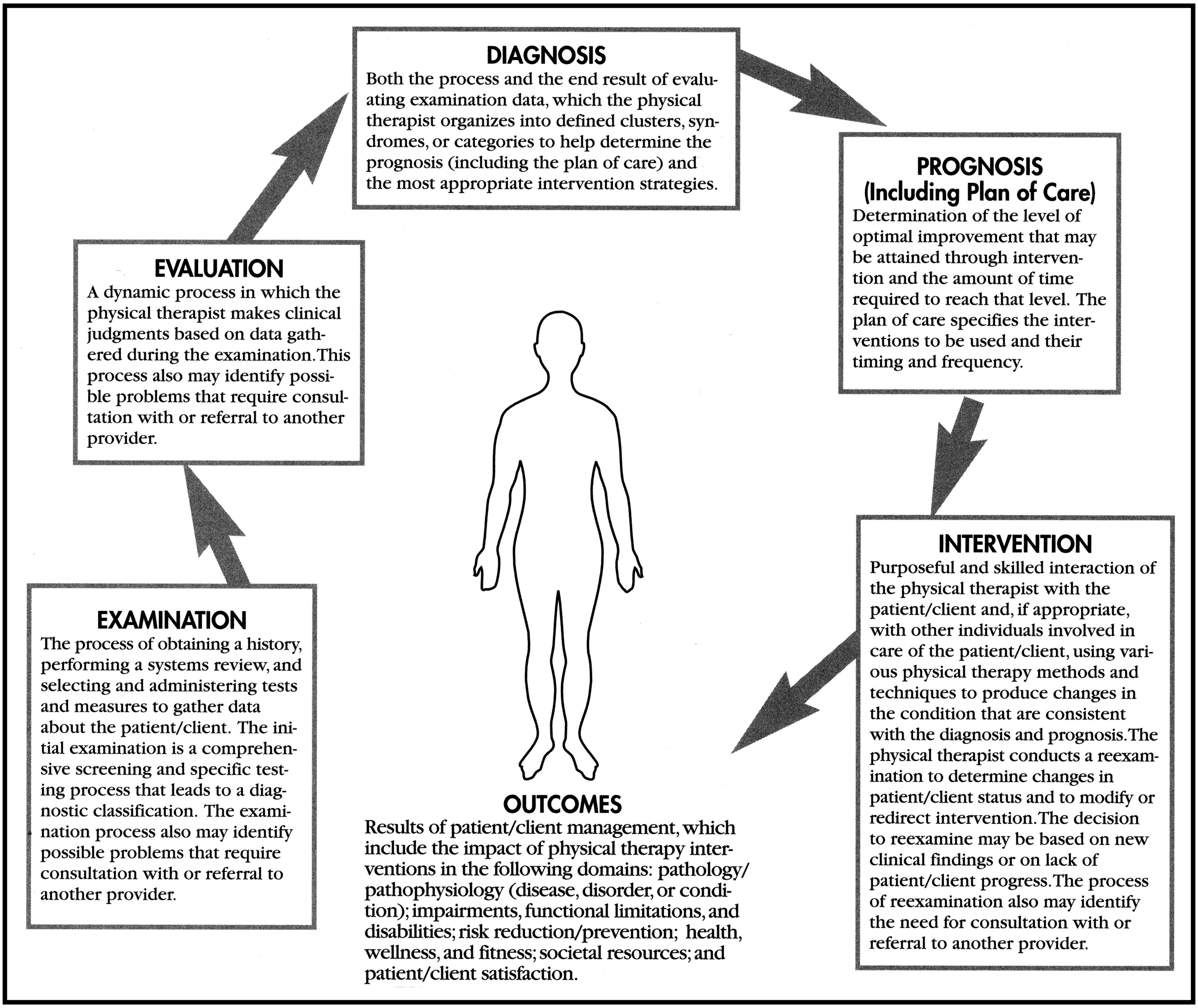
As explicated by the Guide, diagnosis requires gathering of data through examination. During the initial examination, data are obtained through the history, systems review, and selected tests and measures.[3] Therefore, questions of history and the screening procedures performed during the review of systems are also considered diagnostic tests, along with the various tests performed and measurements obtained. Throughout the examination, data are gathered to evaluate and to form clinical judgments. The result of this diagnostic process is a label, or classification, designed to specifically direct treatment. Individual pieces of data are collected for different purposes during the process.[6,7] Some data are collected to focus the examination on a region of the body or to identify a particular pathology (eg, screening tests). Other data are gathered for the purpose of selecting an intervention (eg, tests used for classification). In determining the accuracy of a diagnostic test, the intended purpose of the test should be considered.[8]

JM Fritz, PT, PhD, ATC, is Assistant Professor, Department of Physical Therapy, University of Pittsburgh, 6035 Forbes Tower, Pittsburgh, PA 15260 (USA) (jfritz@pitt.edu). Address all correspondence to Dr Fritz.

RS Wainner, PT, PhD, OCS, ECS, is Physical Therapy Research Coordinator, Wilford Hall Medical Center, Lackland Air Force Base, San Antonio, Tex.

Both authors provided concept/project design and writing.

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Department of the Air Force, the Department of the Army, the Department of the Navy, or the Department of Defense.



DIAGNOSIS
Both the process and the end result of evaluating examination data, which the physical therapist organizes into defined clusters, syndromes, or categories to help determine the prognosis (including the plan of care) and the most appropriate intervention strategies.

PROGNOSIS (Including Plan of Care)
Determination of the level of optimal improvement that may be attained through intervention and the amount of time required to reach that level. The plan of care specifies the interventions to be used and their timing and frequency.

EVALUATION
A dynamic process in which the physical therapist makes clinical judgments based on data gathered during the examination.This process also may identify possible problems that require consultation with or referral to another provider.

INTERVENTION
Purposeful and skilled interaction of the physical therapist with the patient/client and, if appropriate, with other individuals involved in care of the patient/client, using various physical therapy methods and techniques to produce changes in the condition that are consistent with the diagnosis and prognosis.The physical therapist conducts a reexamination to determine changes in patient/client status and to modify or redirect intervention.The decision to reexamine may be based on new clinical findings or on lack of patient/client progress.The process of reexamination also may identify the need for consultation with or referral to another provider.

EXAMINATION
The process of obtaining a history, performing a systems review, and selecting and administering tests and measures to gather data about the patient/client. The initial examination is a comprehensive screening and specific testing process that leads to a diagnostic classification. The examination process also may identify possible problems that require consultation with or referral to another provider.

OUTCOMES
Results of patient/client management, which include the impact of physical therapy interventions in the following domains: pathology/pathophysiology (disease, disorder, or condition); impairments, functional limitations, and disabilities; risk reduction/prevention; health, wellness, and fitness; societal resources; and patient/client satisfaction.

Figure 1.
Five interrelated elements of patient management. (Reprinted with permission of the American Physical Therapy Association from the *Guide to Physical Therapy Practice* [2nd ed].[3])

Although, according to the Guide, the end result of the diagnostic process should most often be a classification grouping based largely on impairments and functional limitations instead of pathoanatomy, individual tests may be used to focus the examination or detect conditions not appropriate for physical therapy management. Tests used in this manner need to demonstrate accuracy for identifying the underlying pathoanatomy. An example of a test used for this purpose is the ankle-arm index,[9,10] a ratio of ankle to arm systolic blood pressure, as a method of screening for atherosclerotic diseases. Some studies have shown that low ankle-arm index values are indicative of various atherosclerotic diseases,[9,10] and such a finding during an examination may indicate the need for referral of the patient to a physician. Another example would occur during the examination of an elderly patient with symptoms in both the lumbar and hip regions. The therapist may want to determine whether the hip symptoms indicate degenerative changes of the hip or whether the symptoms are referred from the lumbar region. Various tests and measures might be considered helpful in making this determination; however, the measurements with the highest diagnostic accuracy for detecting degenerative changes of the hip have been shown to be hip medial (internal) rotation range of motion of less than 15 degrees and hip flexion range of motion of less than 115 degrees.[11] The occurrence of these impairments during the examination, therefore, could provide useful diagnostic information, indicating a need to focus the examination on the hip region.

Some diagnostic tests are performed by physical therapists because the results, singularly or in combination with other findings, are believed to indicate that a particular intervention will be most effective in maximizing the patient's outcome. Tests used in this manner form the foundation of classification systems and need to demonstrate accuracy for identifying which interventions might be useful. For example, the observation of frontal-plane displacement of the shoulders relative to the pelvis (ie, lumbar lateral shift) in a patient with low back pain (LBP) is frequently cited as an important examination finding.[12–16] This finding has been considered by some to be diagnostic of a lumbar disk herniation[16,17]; however, the diagnostic accuracy of a lateral shift for detecting the presence of a disk herniation is poor.[14] Other measures, such as the straight-leg-raise test, serve as more accurate diagnostic tests for the presence of a lumbar disk herniation.[18,19] Despite the lack of accuracy for diagnosing a disk herniation, a lateral shift may be meaningful, not based on its ability to indicate a specific pathoanatomical origin, but because it may indicate which intervention (ie, correction of the lateral shift) will be most useful in reducing pain and disability.[15,20] Although it lacks accuracy for detecting a disk herniation, the presence of a lateral shift may still have diagnostic value if it can be demonstrated that patients judged to have a lateral shift who are treated with correction of the shift have outcomes superior to those of patients treated with alternative approaches. No studies to date have investigated this hypothesis.

In summary, both clinicians and researchers need to consider the purpose for which a diagnostic test is performed. Tests may serve to focus and refine the examination, or they may be used for classification with the goal of selecting effective interventions. The same test may have the potential to serve both purposes, whereas some tests may be useful for one purpose or neither purpose. This distinction is important in considering how to use the diagnostic process in an evidence-based manner. The purpose of a test has important implications for examining the evidence in support of the use of the test and applying the test to clinical practice.

Evidence-Based Practice and the Diagnostic Process

Recently, the term "evidence-based practice" has entered the lexicon of physical therapists, as it has for most medical professionals. *Evidence-based practice* has been defined by proponents as "the conscientious and judicious use of current best evidence in making decisions about the care of individual patients."[21(p71)] Implicit in this definition is the need for a method of determining what constitutes the "best" evidence and how to apply evidence in clinical practice. Substantial effort has gone into the development and dissemination of methods for grading evidence as it relates to treatment effectiveness. Several hierarchical schemes have been promulgated for the purpose of ranking evidence from studies concerning treatment outcomes.[22–24] Although the schemes have some variations, all emphasize the importance of factors such as random assignment to treatment groups, completeness of follow-up, and blinding of examiners and patients in determining the quality of evidence. Although principles for evaluating the quality of an article on treatment outcomes are relatively well known, some authors[8] contend that the question being asked should determine the nature of the evidence to be sought. Therefore, when seeking to answer a diagnostic question, the rules governing the evaluation of studies regarding treatment outcomes are no longer applicable. Rules for judging evidence offered by a study of a diagnostic test have been elucidated; however, they tend to be less widely known and frequently remain unheeded by researchers designing and reporting studies in this area.[25–28] Knowledge of the issues that are important for determining the strength of evidence offered by studies of diagnostic tests is important if the professional dialogue on the diagnostic process in physical therapy is to move forward within a context of evidence-based practice.

Central to the concept of evidence-based practice is the integration of evidence into the management of patients. Integration cannot be reduced to a dichotomy (eg, "use the test or don't use the test") but instead involves a complex interaction between the strength of the evidence offered through use of a test and the unique presentation of an individual patient. Diagnostic tests cannot simply be deemed good or bad. The same test may provide important information for certain patients under certain conditions, but not for others. For example, testing vibration perception is useful for diagnosing a lack of protective sensation and an increased risk of ulceration in the feet of patients with diabetes.[29] However, vibration perception deficits are of more limited diagnostic value in the examination of a patient suspected of having lumbar spinal stenosis.[30]

We will next examine further 2 aspects of evidence-based practice as they apply to the diagnostic process. First, we will discuss 2 of the most important considerations for the evaluation of the strength of evidence related to diagnostic tests: study design and data analysis.[25–27] Second, we will examine the integration of the evidence into the diagnostic process.

Evaluating the Evidence—Study Design

The strength of evidence provided by any study will be substantially affected, and potentially limited, by the study's design. The optimal study design is the one that most effectively reduces susceptibility to bias (ie, a deviation of the results from the truth in a consistent

Table 1.
Contingency Table Created by Comparing the Results of the Diagnostic Test and the Reference Standard

	Reference Standard Positive	Reference Standard Negative
Diagnostic test positive	True positive results A	False positive results B
Diagnostic test negative	C False negative results	D True negative results

Table 2.
Potential Pitfalls for the 3 Most Important Variables Related to the Design of a Study of a Diagnostic Test

Study Variable	Potential Pitfalls
Reference standard	Insufficiently definitive of the condition of interest Not consistent with the intended purpose of the diagnostic test Not applied consistently in all subjects (verification bias) Not independent of the diagnostic test (incorporation bias) Judged by an examiner who is not blinded to the diagnostic test result and clinical condition of the subjects (review bias)
Diagnostic test	Intended purpose of the test not clearly defined Lack of clarity in the description of the test performance Lack of clarity in the description of the test interpretation Judged by an examiner who is not blinded to the results of the reference standard (review bias)
Study population	Study subjects not representative of the population on whom the test is used clinically (spectrum bias)

direction).[27,31] For studies investigating treatment outcomes, the design best accomplishing this objective is recognized as the randomized clinical trial. However, if the research question is one of diagnosis, the randomized trial is no longer the most desirable design. The optimal design for examining a diagnostic test, in the opinion of experts, is "a prospective, blind comparison of the test and the reference test in a consecutive series of patients from a relevant clinical population."[26(p1062)] That is, a study investigating a diagnostic test should utilize a prospective cohort design in which all subjects are evaluated using the diagnostic test or tests and a reference standard representing the definitive, or best, criteria for the condition of interest. When performed in this manner, the results of the test and the reference standard can be summarized in a 2×2 table, as depicted in Table 1.

Issues beyond the basic design of a study are important for determining the extent to which the potential for bias has been minimized in a study and for determining the strength of the evidence. For studies of diagnostic tests, the most important issues are the reference standard, the diagnostic test, and the population studied. The most important considerations for each issue are summarized in Table 2 and are described below.

The Reference Standard

In a study of a diagnostic test, the test of interest is compared with a reference standard. The reference standard is the criterion that best defines the condition of interest.[32] For example, if a test is performed to determine the presence of a meniscal tear in the knee, the most appropriate reference standard would be observation of the meniscus with arthroscopy. The reference standard should have demonstrated validity that justifies its use as a criterion measurement.[33] If the reference standard is determined to lack validity, little meaningful information can be derived from the comparison.[34] The validity of the reference standard may be compromised by several factors.

First, the reference standard should possess acceptable measurement characteristics, as defined by the APTA's standards for tests and measurements.[33,35] For example, the Ashworth scale has become a commonly used measure of "muscle spasticity."[36] Despite several studies questioning the reliability and construct validity of measurements obtained with the scale in either its original or modified form,[37–40] the Ashworth scale continues to be used as a reference standard.[41] If a reference standard is not reproducible, or lacks a strong conceptual basis for its use, it should not be used as the criterion against which to judge the adequacy of another test.[26]

The reference standard should also be consistent with the intended purpose of the diagnostic test. The majority of reference standards used to study diagnostic tests have been measures of pathoanatomy.[42] If a pathoanatomical reference standard is consistent with the test's purpose, this could serve as a valid measure for comparison. If a diagnostic test is used to select interventions with the goal of maximizing outcomes, a measure of pathoanatomy is unlikely to serve as an appropriate reference standard. As defined by the Guide,[3] outcomes are measures of functional limitations, disability, patient satisfaction, and prevention; therefore, diagnostic tests used to select interventions should be tested against a reference standard related to one of these measures.

An investigation by Burke et al[43] of the Phalen test, which is commonly used for patients with suspected carpal tunnel syndrome (CTS), provides an example of selecting reference standards consistent with the purpose of the diagnostic test. The Phalen test could be examined as a screening test to detect compression of the median nerve or as a test indicating the need for specific interventions (eg, wrist splinting).[43] To reflect these different purposes, Burke et al[43] compared the Phalen test against 2 reference standards: results of a nerve conduction velocity study and patient-reported improvement after a 2-week course of wrist splinting. The nerve conduction study in which the distal motor and sensory latencies of the median nerve were measured served as the reference standard for an examination of the Phalen test's ability to detect nerve compression. Patient-reported improvement after 2 weeks served as a reference standard for the accuracy of the Phalen test as an indication of whether wrist splinting was useful as an intervention.

If the reference standard is not consistent with the purpose of the diagnostic test, the results become difficult to interpret. For example, 2 recent studies examined various tests for sacroiliac (SI) region dysfunction in patients receiving physical therapy.[44,45] In both studies, the reference standard was the presence of LBP, judged as positive (patient consulting for LBP) or negative (patient consulting for an upper-extremity condition). By using this reference standard, the researchers examined the accuracy of the SI region tests in distinguishing between individuals with and without LBP. It does not appear, however, that this reference standard is consistent with the purpose of these tests. In the literature, SI region tests are proposed to distinguish patients thought to have SI region dysfunction from those with LBP related to other syndromes[46–48] or to determine whether a patient is likely to respond to a particular intervention designed for SI region dysfunction (eg, SI region manipulation).[12,49] The results of studies using a reference standard of the presence of LBP when the issue is whether there is SI region dysfunction are difficult to interpret because this standard is inconsistent with the purposes for which the tests are commonly used. Determining the usefulness of the tests based on the results of these studies may lead to erroneous conclusions.

Improper use of reference standards in a study may compromise the validity of the research. The reference standard should be applied consistently to all subjects.[25,26,32] If the reference standard is expensive or difficult to obtain, it may not be performed on subjects with a low probability of having the condition. Verification (or workup) bias occurs when not all subjects are assessed by use of the reference standard in the same way.[27,50] A common example of verification bias is demonstrated by a study of diagnostic accuracy of tests for posterior cruciate ligament (PCL) integrity.[51] The reference standard was magnetic resonance imaging (MRI), an appropriate pathoanatomical reference standard for PCL integrity. A group of individuals with no history of knee injury were included in the study. These individuals were assumed to have an intact PCL without MRI verification.[51] Another example comes from a study of a screening examination using goniometry for detecting cerebral palsy in preterm infants.[52] Goniometric measurements were taken at the hip, knee, and ankle. If the range of motion measurements fell outside a normal range, the child was believe to be at an increased risk of having cerebral palsy.[52] Infants with a high suspicion of cerebral palsy were referred to a neurologist whose evaluation then served as the reference standard. Only 97 of 721 infants were referred, and a less rigorous reference standard consisting of chart reviews was used for the remaining subjects.[52] The impact of verification bias is related to the likelihood that an individual not assessed with the reference standard could have the condition. It may be unlikely that an individual with no history of knee injury would have a compromised PCL. The adequacy of using a chart review for identifying cerebral palsy may leave this study more susceptible to verification bias. Verification bias can lead to an overestimation of diagnostic accuracy.[26,53]

The reference standard should also be independent of the diagnostic test. Incorporation bias occurs when the reference standard includes the diagnostic test being studied.[54] An example comes from a study of single-leg hop tests for diagnosing anterior cruciate ligament (ACL) integrity.[55] The authors evaluated 50 subjects with a chronic ACL-deficient knee and 60 subjects with no prior knee injury. All subjects performed the hop tests. The reference standard was defined as the mean (±2 standard deviations) of the absolute value of the right-to-left difference in time to complete the test in the subjects without knee injury. The authors then applied this standard to the results of all 110 subjects and found high levels of diagnostic accuracy for the tests in distinguishing the 2 groups of subjects.[55] This result is not surprising given that the interpretation of the reference standard was based on the test results of the subjects without knee injury. Incorporation bias is also likely to inflate the accuracy of a diagnostic test.[26]

The reference standard should be judged by an individual who does not know the diagnostic test results and the overall clinical presentation of the subject.[26,53,56] If blinding is not maintained, judgments of the reference standard may be influenced by expectations based on knowledge of the test results or by some other clinical information.[56] Review bias may occur if either the reference standard or the diagnostic test is judged by an individual with knowledge of the other result.[53]

The Diagnostic Test

Practitioners and researchers should be able to describe diagnostic tests in sufficient detail to permit replication of the tests by other therapists. We contend that test descriptions should cover 3 aspects: the intended use, physical performance, and scoring criteria. The intended clinical use of a test is an important consideration, although this aspect of the test description is often overlooked by researchers and practitioners.[25] As indicated previously, a diagnostic test may be used for a variety of purposes. If researchers do not clarify the intended purpose of a test under study, it is difficult to assess the appropriateness of the reference standard. When clinicians do not consider the purpose of diagnostic tests used in practice, they are susceptible to viewing tests as either good or bad, without recognition that a test may be useful for one purpose, but inappropriate for another purpose. For example, the KT-1000 knee arthrometer* possesses a high degree of diagnostic accuracy for distinguishing between individuals with and without ACL deficiency,[57,58] but it has not been shown to be useful for assisting in the selection of an intervention (surgical versus nonsurgical).[59]

The manner in which a test is performed should be detailed. A study's results can be generalized to a clinical setting only if a test is performed as it was performed in the study. For example, Katz and Fingeroth[60] compared various tests for ACL integrity against a reference standard of observation of the ligament during arthroscopy. The Lachman test demonstrated very good diagnostic accuracy for ACL integrity; however, the test was performed with the subjects under anesthesia. If the results were accepted without consideration of the manner in which the test was performed, a clinician may have unrealistic expectations of the usefulness of test results when applying the test to patients who are not under anesthesia. This is illustrated by a study of the Lachman test performed by physical therapists in a clinical setting that led to lower levels of diagnostic accuracy.[61]

The description of a diagnostic test should include the criteria used to determine positive and negative results. Many tests used in physical therapy, though well known, may have varied or unclear grading criteria. Testing for centralization in patients with LBP is an example. There is general agreement that centralization is an important diagnostic finding,[62–64] but no such consensus exists on precisely what constitutes centralization. Some therapists use definitions strictly based on movement of symptoms from distal to proximal,[16,65] whereas other therapists define centralization to include diminishment of pain during testing.[63] Such disagreements are not unique to judgments of centralization, and it is crucial for authors to clarify how they defined positive and negative results. It is also important to indicate whether the test cannot be performed or the results are indeterminate for any subjects. Because these occurrences could influence the clinical use of a test, they should be reported and explained.[53,66] Measurements obtained with a test also are susceptible to review bias, as previously explained. Review bias can be avoided if the measurements and judgments are done by individuals who are blinded to the reference standard. Diagnostic accuracy may be overestimated if blinding is not maintained.[26]

The Study Population

Subjects included in a study of a diagnostic test should consist of individuals who would be likely to undergo the test in clinical practice.[26,53] This also means that individuals who are positive on the reference standard should reflect a continuum of severity, from mild to severe, whereas those who are negative with respect to the reference standard should have conditions commonly confused with the condition of interest and should not be a group of control subjects without impairments or disabilities.[34] Many of the tests already cited in this perspective have used groups of subjects without impairments or disabilities who were chosen out of convenience. When subjects without any symptoms, impairments, or disabilities are tested, this does not reflect the way most tests are applied clinically, where distinctions between individuals with similar symptoms are required. Any test should at least be expected to demonstrate greater diagnostic accuracy when attempting to distinguish between individuals without symptoms and those with severe conditions.[56] Spectrum (or selection) bias may occur when study subjects are not representative of the population on whom the test is typically applied in practice.[26] Spectrum bias, in our opinion, can profoundly affect the results of a study.[26]

The best method of ensuring a representative sample and avoiding spectrum bias is to utilize a prospective cohort design with a consecutive group of subjects from a clinical population. Use of a case-control design with retrospective selection of subjects for inclusion makes a study susceptible to spectrum bias.[53] This type of design occurs when a group of subjects with the condition of interest and a group of comparison subjects are assembled for examination. Even if the use of subjects without known impairments or disabilities is avoided, case-control designs can distort the typical mix of subjects seen in a clinical setting by artificially controlling the prevalence and presentation of the condition of interest, potentially affecting the accuracy and utility of a diagnostic test.[28,54,67,68]

A comparison of studies examining the diagnostic accuracy of the Phalen test for detecting median nerve compression in the carpal tunnel provides an example of the impact of spectrum bias (Tab. 3). The study by Burke et al[43] and 2 other studies[69,70] compared the Phalen test against a reference standard involving nerve conduction

* MEDmetric Corp, 7542 Trade St, San Diego, CA 92121.

Table 3.
Comparison of Studies Examining the Accuracy of the Phalen Test for Diagnosing Compression of the Median Nerve Within the Carpal Tunnel[a]

	Kulhman and Hennessey[70]	Burke et al[43]	Gellman et al[69]
Reference standard	Nerve conduction study (any one of the following): 1. Median motor onset distal latency ≥1.0 ms longer than ulnar motor onset distal latency 2. Median sensory peak distal latency to the thumb ≥0.5 ms longer than radial sensory peak distal latency to the thumb 3. Median sensory peak distal latency to the long finger ≥0.5 ms longer than ulnar sensory peak distal latency to the small finger	Nerve conduction study (any one of the following): 1. Minimum median sensory distal latency measured at 14 cm of 4.1 ms and a minimum motor distal latency at 8 cm of 4.4 ms 2. Median sensory distal latency >0.5 ms longer than ulnar sensory distal latency	Nerve conduction study (any one of the following): 1. Minimum median sensory distal latency of >3.5 ms, or of 1 ms more than the opposite side 2. Minimum median motor distal latency of 4.5 ms, or of 1 ms more than the opposite side
Diagnostic test performance	The subject actively places the wrists in complete, but unforced, flexion for 60 s	Not described	The subject actively places the wrist in complete, but unforced, flexion for 60 s
Diagnostic test grading	If numbness or paresthesia is produced or exaggerated in the hand, the test is positive	Not described	If numbness and tingling are produced or exaggerated in the median nerve distribution of the hand, the test is positive
Study population	180 consecutive subjects (228 hands) referred for electrodiagnostic consultation with suspected CTS	186 subjects (290 hands) referred for splinting with a history consistent with CTS	106 hands with symptoms consistent with CTS, 16 hands with symptoms commonly confused with CTS, 50 asymptomatic hands
Sensitivity (%) (95% CI)	51 (43, 60)	51 (44, 58)	71 (59, 81)
Specificity (%) (95% CI)	76 (66, 83)	54 (35, 71)	80 (67, 89)
Overall accuracy (%)	60	52	72

[a] Accuracy represents the percentage of correct results on the diagnostic test when compared with the reference standard. CTS=carpal tunnel syndrome, CI=confidence interval.

velocity studies. Similar criteria for judging the reference standard were used in the 3 studies. The description of how the diagnostic test was performed and the grading criteria were nearly identical in 2 studies, but they were not reported in the third study. The greatest difference among the studies was the subjects. In 2 studies,[43,70] there were cohorts of subjects with symptoms consistent with CTS. In the third study,[69] subjects included those with symptoms consistent with CTS, a few with known diagnoses other than CTS but with a similar presentation (eg, diabetic peripheral neuropathy), and 25 subjects (50 hands tested) without symptoms consistent with CTS. Inclusion of people without symptoms creates a spectrum bias by assembling a population unrepresentative of the clinical population in which the test is typically used. As would be anticipated, the study most subject to spectrum bias also demonstrated the highest level of diagnostic accuracy for the Phalen test (Tab. 3).

Evaluating the Evidence—Data Analysis

The basic layout for the data analysis in a study of a diagnostic test is depicted in Table 1. The result for each subject fits into only 1 of the 4 categories based on a comparison of the results of the diagnostic test and the diagnosis based on the reference standard. Results in categories "a" (true positive) and "d" (true negative) represent correct test results, whereas categories "b" (false positive) and "c" (false negative) contain erroneous results. From this basic layout, several statistics can be calculated (Tab. 4).[56]

The overall accuracy of a test can be determined by dividing the number of correct results by the total number of tests conducted.[56] A perfect test would have an overall accuracy of 100%; however, no test used in clinical practice can be expected to demonstrate this level of accuracy, and the goal is to characterize the nature of the errors.[71] The overall accuracy of a test does not distinguish between false positive and false negative results and therefore has limited usefulness.[72]

Sensitivity, Specificity, and Predictive Values

Sensitivity and specificity values are calculated vertically from the 2 × 2 table and represent the proportion of correct test results among individuals with and without the condition, respectively. *Sensitivity* (or true positive

Table 4.
Statistics Commonly Used to Examine Diagnostic Tests

Statistic	Formula	Description
Overall accuracy	(a + d)/(a + b + c + d)	The proportion of test results that are correct
Positive predictive value	1/(a + b)	Given a positive test result, the probability that the individual has the condition
Negative predictive value	d/(c + d)	Given a negative test result, the probability that the individual does not have the condition
Sensitivity	a/(a + c)	Given that the individual has the condition, the probability that the test will be positive
Specificity	d/(b + d)	Given that the individual does not have the condition, the probability that the test will be negative
Positive likelihood ratio	sensitivity/(1 − specificity)	Given a positive test result, the increase in odds favoring the condition
Negative likelihood ratio	(1 − sensitivity)/specificity	Given a negative test result, the decrease in odds favoring the condition

Table 5.
Accuracy of Weakness of the Extensor Hallucis Longus Muscle for Diagnosing L5 Radiculopathy in the Study by Lauder et al[75,a]

	L5 Radiculopathy Present	L5 Radiculopathy Not Present	
Weakness positive	6 A	38 B	Positive predictive value: (6/46)=.14, 95% CI: .06, .27
Weakness negative	C 4 Sensitivity (%): (6/10)=.60 95% CI: .31, .83	D 46 Specificity (%): (46/84)=.55 95% CI: .44, .65	Negative predictive value: (46/50)=.92, 95% CI: .81, .97

[a] Prevalence of L5 radiculopathy in this study was 11%. CI=confidence interval.

rate) is the proportion of subjects with the condition who have a positive test result. *Specificity* (or true negative rate) is the proportion of subjects without the condition who have a negative test result.[42]

Predictive values are calculated horizontally from the 2 × 2 table and represent the proportion of subjects with a positive or negative test result that are correct results. The *positive predictive value* is the proportion of subjects with a positive test result who actually have the condition. The *negative predictive value* is the proportion of subjects with a negative test result who do not have the condition.[73]

Predictive values might appear to be more useful for applying the results of a study because these values relate to the way these tests are used in clinical decision making: given a test result (positive or negative), what is the probability that the result is correct? Sensitivity and specificity values work in the opposite direction: given the condition is present or absent, what is the probability that the correct test result will be obtained? Despite their apparent usefulness, predictive values can be deceptive because they are highly dependent on the prevalence of the condition of interest in the sample. Positive predictive values will be lower and negative predictive values will be higher in samples with a low prevalence of the condition. If prevalence is high, the trends reverse.[74] Sensitivity and specificity values remain fairly consistent across different prevalence levels.[42] A comparison of 2 studies examining the diagnostic accuracy of weakness of the extensor hallucis longus muscle for detecting L5 radiculopathy illustrates this point. Lauder et al[75] studied consecutive patients referred to physical medicine physicians with a suspicion of lumbar radiculopathy (Tab. 5). The reference standard was electromyographic findings, and, based on this standard, the prevalence of L5 radiculopathy was 11% (10/94). Kortelainen et al[76] studied patients referred for surgery with symptoms of sciatica (Tab. 6). Based on a reference standard of surgical observation of the nerve root, the prevalence of L5 radiculopathy was 57% (229/403). The sensitivity and specificity values remained fairly consistent. The predictive values, however, varied greatly between studies due to disparate prevalence rates, with the study with higher prevalence of radiculopathy showing a higher positive predictive value.

Sensitivity and specificity values provide useful information for interpreting the results of diagnostic tests. Sensitivity represents the ability of the test to recognize the condition when present. A highly sensitive test has relatively few false negative results. High test sensitivity, therefore, attests to the value of a negative test result.[77,78] Sackett et al[42] have advocated using the acronym

Table 6.
Accuracy of Weakness of the Extensor Hallucis Longus Muscle for Diagnosing L5 Radiculopathy in the Study by Kortelainen et al[76,a]

	L5 Radiculopathy Present	L5 Radiculopathy Not Present	
Weakness positive	126 A	54 B	Positive predictive value: (126/180)=.70, 95% CI: .63, .76
Weakness negative	C 103 Sensitivity (%): (126/229)=.55 95% CI: .49, .61	D 120 Specificity (%): (120/174)=.69 95% CI: .62, .75	Negative predictive value: (120/223)=.54, 95% CI: .47, .60

[a] Prevalence of L5 radiculopathy in this study was 57%. CI=confidence interval.

Table 7.
Accuracy Statistics of Clinical Tests for Diagnosing Subacromial Impingement Syndrome[a]

Test	Sensitivity (%)	Specificity (%)	Positive LR	Negative LR
Hawkin	92 (84, 96)	25 (14, 42)	1.2 (1.0, 1.5)	0.32 (0.12, 0.76)
Neer	89 (80, 94)	31 (17, 46)	1.3 (1.0, 1.6)	0.37 (0.18, 0.86)
Horizontal adduction	82 (73, 89)	28 (15, 43)	1.1 (0.90, 1.4)	0.65 (0.32, 1.4)
Speed	69 (58, 77)	56 (40, 71)	1.5 (1.0, 2.3)	0.57 (0.37, 0.87)
Yergason	37 (28, 48)	86 (70, 94)	2.7 (1.1, 6.0)	0.73 (0.59, 0.91)
Painful arc	33 (23, 43)	81 (63, 90)	1.7 (0.76, 3.3)	0.84 (0.68, 1.1)
Drop arm	8 (4, 16)	97 (85, 100)	2.8 (0.35, 21.7)	0.95 (0.87, 1.3)

[a] Numbers in parentheses represent 95% confidence intervals, which were estimated from the data presented in the study.[79] LR=likelihood ratio.

"SnNout" (if sensitivity [Sn] is high, a negative [N] result is useful for ruling out [out] the condition). High sensitivity indicates that a test can be used for excluding, or ruling out, a condition when it is negative, but does not address the value of a positive test. Specificity indicates the ability to use a test to recognize when the condition is absent. A highly specific test has relatively few false positive results, and therefore speaks to the value of a positive test.[77,78] The acronym applicable in this case is "SpPin" (if specificity [Sp] is high, a positive [P] result is useful for ruling in [in] the condition).[42]

Unfortunately, few tests possess both high sensitivity and specificity. Knowledge of the sensitivity and specificity of a test can help clinicians refine clinical decision making by allowing them to weigh the relative value of positive or negative results. A recent study[79] examining the diagnostic accuracy of clinical tests for detecting subacromial impingement syndrome provides an example. Six tests were compared against a reference standard of MRI of the supraspinatus tendon. No test had high levels of both sensitivity and specificity (Tab. 7). The Hawkin test was the most sensitive, and the drop arm test was most specific.[79] The high sensitivity (92%) indicates that a negative Hawkin test is useful for ruling out subacromial impingement. The low specificity (25%), however, signifies that a positive Hawkin test has little meaning. The drop arm test was very specific (97%), indicating that a positive test is useful for confirming subacromial impingement. The sensitivity of the drop arm test was poor (8%), revealing a high number of false negative results and attesting to the lack of meaning of a negative result.

Likelihood Ratios

Sensitivity and specificity values provide useful information; however, they have several shortcomings. These values work in the opposite direction of clinical decision making. Clinicians have knowledge of the test result and want to infer the probability that the result is correct. Sensitivity and specificity values infer the probability of a correct test, given the result of the reference standard. Sensitivity and specificity values can be used as independent estimates of the usefulness of negative and positive test results, but this information cannot be combined and analyzed simultaneously. The actual performance of a diagnostic test is not only related to sensitivity and specificity values, but also dependent on the pretest probability that the condition is present. Useful tests should produce large shifts in probability once the result of the test is known.[77,80,81] Sensitivity and specificity values cannot be used to quantify the shift in probability of the condition given a certain test result.

The best statistics for summarizing the usefulness of a diagnostic test are likelihood ratios.[82,83] Likelihood ratios (LRs) overcome the difficulties cited by reflecting a combination of the information contained in sensitivity and specificity values into a ratio that can be used to quantify shifts in probability once the diagnostic test results are known.[84] The positive LR is calculated as sensitivity/(1 − specificity) and indicates the increase in

Table 8.
A Guide to Interpretation of Likelihood Ratio (LR) Values[a]

Positive LR	Negative LR	Interpretation
>10	<0.1	Generate large and often conclusive shifts in probability
5–10	0.1–0.2	Generate moderate shifts in probability
2–5	0.2–0.5	Generate small, but sometimes important, shifts in probability
1–2	0.5–1	Alter probability to a small, and rarely important, degree

[a] Adapted from Jaeschke et al.[83]

odds favoring the condition given a positive test result. The negative LR is calculated as (1 − sensitivity)/specificity and indicates the change in odds favoring the condition given a negative test result.[27] An LR of 1 indicates that the test result does nothing to change the odds favoring the condition, whereas an LR greater than 1 increases the odds of the condition, and an LR less than 1 diminishes the odds of the condition. Table 8 provides a guide for interpreting the strength of an LR.[83]

A positive LR indicates the shift in odds favoring the condition when the test is positive. It is desirable, therefore, to have a large positive LR. Tests with a large positive LR generally have high specificity because both values attest to the usefulness of a positive test. In the study by Calis et al,[79] for example, the drop arm test had the highest specificity (97%) for determining the presence of subacromial impingement syndrome and also the largest positive LR (2.8) (Tab. 7). Because the negative LR indicates the change in odds favoring the condition given a negative result, a small negative LR will indicate a test that is useful for ruling out a condition when negative. Small negative LR values correspond to high sensitivity, as illustrated by the subacromial impingement syndrome tests. The highest sensitivity and smallest negative LR were found for the Hawkin test. A comparison of the horizontal adduction and Speed tests indicates the importance of combining sensitivity and specificity values. The sensitivity of the Speed test (69%) was less than that of the horizontal adduction test (82%). However, because the Speed test was substantially more specific than the horizontal adduction test (56% versus 28%), the negative LR was smaller for the Speed test (0.57 versus 0.65).

Diagnostic tests measured on a continuous scale are frequently transformed into multilevel ordinal outcomes based on cutoff scores. When this is the case, LR values can be calculated for each level of the test.[42] Riddle and Stratford[85] illustrated this process using the Berg Balance Test. Different test results were used as cutoff scores, and the LR values were calculated for each level. A more detailed explanation of the process can be obtained from the article by Riddle and Stratford.[85]

Evaluating the Evidence—Additional Considerations

Confidence Intervals

As is true of all statistics, sensitivity, specificity, and LR values are taken from a sample and represent an estimate of the true value that could be found in the population.[84] The confidence interval (CI) attests to the precision of this estimate. A 95% CI is the most common and indicates a range of values within which the population value would lie with 95% certainty.[86] If the CI is wide and contains values that are not clinically important, the usefulness of the measure may be questionable. That is, if another estimate were taken from a different sample, the statistic calculated might be substantially different. In the study by Calis et al,[79] for example, the drop arm test had the largest positive LR among the tests for subacromial impingement (2.8), but the 95% CI was wide (0.35–21.7), indicating that the positive LR estimated from this sample of 120 patients was not very precise. Formulas for calculating CI ranges for diagnostic statistics have been published.[84,86,87] As is apparent in Table 7, the recommended formulas do not result in a symmetrical CI about the statistical estimate.[88] The asymmetry is more pronounced as the sensitivity and specificity values move farther from 50% in either direction.[86] The width of the CI will also be related to the sample size and the amount of variability in the test being studied. Reporting of a CI with any diagnostic statistic is recommended to permit an assessment of the precision of any estimate of diagnostic accuracy.[86,89]

The Chi-Square Statistic

Studies of diagnostic tests comparing categorical results of a test and a reference standard are frequently analyzed with a chi-square statistic and accompanying significance level. The chi-square statistic tests the hypothesis that the test results and reference standard have no association, but it does not indicate the strength or direction of any relationship that exists.[90] Chi-square statistics and associated probability values cannot assist in the process of probability revision based on test results in individual patients and, therefore, cannot be considered evidence-based statistics.[91]

Conclusions based strictly on chi-square analyzes can be misleading without information on sensitivity, specificity, and LR values. The study by Burke et al[43] on diagnostic tests for patients with suspected CTS illustrates this concern. One diagnostic test examined by the authors was the patient self-report of hand swelling, graded as present (positive) or absent (negative), against a reference standard of response to 2 weeks of splinting. The reference standard was graded as "positive response to splinting" or "no response to splinting" based on patient self-report.[43] The authors chose to analyze the data using

Table 9.
Comparison of the Results of the Patients' Complaints of Hand Swelling and Results of a 2-Week Period of Splinting in a Group of Patients With Suspected Carpal Tunnel Syndrome[43,a]

	Positive Response to Splinting	Negative Response to Splinting
Complaint of swelling positive	17 A	120 B
Complaint of swelling negative	C 34	D 119
	Sensitivity (%)=33.3 95% CI: 20.4, 46.3	Specificity (%)=49.8 95% CI: 43.5, 56.1
χ^2=4.80, *P*=.028		
Positive likelihood ratio=0.66, 95% CI: 0.44, 1.0		
Negative likelihood ratio=1.34, 95% CI: 1.06, 1.69		

[a] The chi-square test shows statistical significance, but the likelihood ratio values indicate a lack of accuracy for the complaint of hand swelling in diagnosing a positive response to splinting. CI=confidence interval.

a chi-square test only and found a statistically significant result (*P*=.028) (Tab. 9). The authors concluded, "These data suggest that the complaint of subjective swelling in the hand or wrist may be one of the most important findings from the history and clinical examination for determining which patients will, in fact, respond to conservative treatment (splinting)."[43] The sensitivity, specificity, and LR values calculated from the data do not support this conclusion. The sensitivity (33.3) and specificity (49.8) were low, resulting in a positive LR of 0.66 and negative LR of 1.34 (Tab. 9). Both LR values are close to 1, with the negative LR slightly greater than 1 and the positive LR slightly less than 1, indicating that the weak relationship between a complaint of swelling and response to splinting is in an inverse direction (ie, a negative complaint of swelling is associated with an increased likelihood of response to splinting). Because evidence-based statistics were not reported, we believe that the authors overinterpreted the utility of the test. This example illustrates the necessity of reporting sensitivity, specificity, and LR values to permit an appropriate assessment of a diagnostic test and interpretation for individual patient decision making.

The Role of Reliability

In order to provide useful information, a test should yield reliable results in the clinical setting. That is, performance of the test on different occasions should yield the same result if the status of the patient being examined has not changed. Traditionally, reliability has been emphasized as a precursor to validity, a preliminary step that should be completed prior to initiating any study of validity. The numerous studies examining diagnostic test reliability without any assessment of validity attest to this mind-set. The peril in this approach is that it may lead to the dismissal of potentially useful tests based on an inability to reach an arbitrary threshold of reliability. This could be due to properties of the statistics used to measure reliability.

The kappa statistic is the reliability coefficient typically used in studies of agreement between examiners for categorical data.[92] The kappa statistics appropriate for this purpose because it is a chance-corrected measure of agreement; however, it can be subject to deflation based on the prevalence of the condition being measured.[35,93] For example, Spitznagel and Helzer[94] noted that, if 2 raters of equal ability each performed a test and each rater was known to have 80% sensitivity and 98% specificity when his or her results were compared with a reference standard, the kappa statistic between the raters would be .67 if the errors made by the raters relative to the reference standard were independent. If the same raters, with the same level of accuracy, repeated the test in a second population with a prevalence of only 5%, the kappa value would fall to .52.[94] This is an example of the difficulty in interpreting kappa values when prevalence is extremely high or low. Many conditions of interest in physical therapy are rare, and kappa statistics used in these instances may be artificially lowered.

In addition, although arbitrary scales exist for categorizing kappa values as poor, fair, good, and so on,[92] the threshold level making a test "reliable enough" is not known. For example, Smieja et al[95] examined the reliability and diagnostic accuracy of tests used in the identification of patients with diabetes who lacked sufficient protective sensation of the feet. A total of 304 patients were examined, 200 of whom were also examined by a second rater to measure reliability. The reference standard was a Semmes-Weinstein monofilament examination. One diagnostic test that was examined was position sense assessed at the interphalangeal joint of the great toe for a 10-degree change. The kappa value between raters for judgments of position sense was only fair by most standards (κ=.28). The results (Tab. 10), however, show that the position sense test provided useful information when it was positive (specificity=98%, positive LR=12.8).[95] If the reliability assessment had been performed separate from the study of validity, it is possible that the position sense test would have been discarded from further consideration due to a lack of reliability, and the potential diagnostic value of a positive result may not have been uncovered.

Reliability data certainly convey meaningful information; however, we believe that their usefulness is best appreciated when considered in conjunction with data examining diagnostic accuracy or utility. Reliability assessments conducted as independent preliminary studies can lead to the premature exclusion of useful tests or the promotion of highly reliable, but diagnostically

Table 10.
Accuracy of Position Sense Testing for Diagnosing a Lack of Protective Sensation in the Feet of Patients With Diabetes[95,a]

	Monofilament Test Positive	Monofilament Test Negative
Position sense test positive	34 A	2 B
Position sense test negative	C 135	D 126
	Sensitivity (%)=20.1 95% CI: 14.1, 26.2	Specificity (%)=98.4 95% CI: 96.3, 1.0
Positive likelihood ratio=12.8, 95% CI: 3.1, 52.2		
Negative likelihood ratio=0.81, 95% CI: 0.75, 0.88		

[a] CI=confidence interval.

meaningless, tests. To encourage complete examination of a diagnostic test, reliability data should be considered a complement to, not a precursor of, an assessment of diagnostic value. An important role of reliability data in the context of assessing the strength of evidence provided by a diagnostic test is that it may provide an explanation for inadequate accuracy or utility.[56,77] When a measurement is found to have little diagnostic meaning and poor reliability, the test's diagnostic ability may be improved if the test is performed in a manner that leads to more reliable measurements.

Applying the Evidence—Practicing Evidence-Based Practice

Although it may not be viewed in this manner by all therapists, the diagnostic process is essentially an exercise in probability revision (Fig. 2).[96] Prior to performing a test, a therapist has some idea of the likelihood that the patient has the condition of interest. The likelihood may be most readily expressed in qualitative terms such as "highly likely," "very unlikely," and so forth. These terms, however, can be made more quantitative by speaking in terms of probabilities. For instance, if a condition is thought to be highly likely, this may translate in the therapist's mind to a probability of 75% or 80% certainty. The condition of interest may be a question of screening (Does the patient's problem involve a certain anatomical structure or region?) or of classification (Is the patient going to respond to a certain treatment?). The therapist can also have in mind a treatment threshold level of certainty at which he or she will be "sure enough" and ready to act.[81] For example, a therapist may feel that he or she must be at least 80% certain that a patient has lumbar spinal stenosis before initiating a program of flexion exercises. Treatment thresholds may not be explicitly stated, but we believe that all therapists reach a point when the examination and evaluation process stops and intervention begins. This threshold should take into consideration the costs associated with being wrong versus the benefits of being correct.[97,98] For example, a high threshold is required when ruling out metastatic disease as a source of LBP. Conversely, if the question concerned the application of a treatment with minimal cost and low potential for side effects, the threshold would be lower. For example, the application of patellar taping for a patient with patellofemoral joint pain is a low-cost intervention with few side effects. A therapist may feel it necessary to be only 50% certain that the treatment will be effective in order to initiate the treatment.

The patient's values should also be considered in establishing treatment thresholds and determining when to implement an intervention.[99] As an example, during the examination of a patient who had a stroke over 1 year previously, a therapist may test the modality of light touch by alternately touching both of the patient's hands and checking for any difference in feeling. If the light touch test is positive (ie, there is a difference in feeling), there is evidence to suggest that the patient has a higher probability of improving function of the hemiplegic upper extremity with an intervention involving forced-use therapy.[100] This intervention, however, requires the patient to immobilize the healthy upper extremity for up to 12 hours per day and attend daily therapy sessions lasting for 6 hours.[100] Some patients may not value the potential increased function of the extremity highly enough to tolerate the required treatment intensity unless the probability of improving function is very high.

The amount of data required to move beyond the treatment threshold is partly determined by the pretest probability that the condition of interest is present. The pretest probability is an important consideration for examining the diagnostic process because it determines how much data will be required to reach a treatment threshold. If the pretest probability that a condition is present is very high, perhaps 80%, one negative test result is unlikely to lower the probability sufficiently to permit its exclusion from further consideration, and additional testing will likely be required to reach a threshold at which the diagnosis would be sufficiently ruled out.[101] Likewise, if the pretest probability is low, a single positive finding will probably not be adequate to elevate the probability beyond the threshold to rule in the condition. That is, if the therapist is fairly certain regarding a diagnosis and an unexpected finding occurs, further data are probably required before a treatment threshold can be reached. Pretest probabilities can come from a variety of sources, including epidemiological data on prevalence rates for certain conditions, information already obtained on the patient from the examination, and clinical experience with similar presentations.[72] Regardless of the source, an often overlooked step in examining the diagnostic process is recognizing and

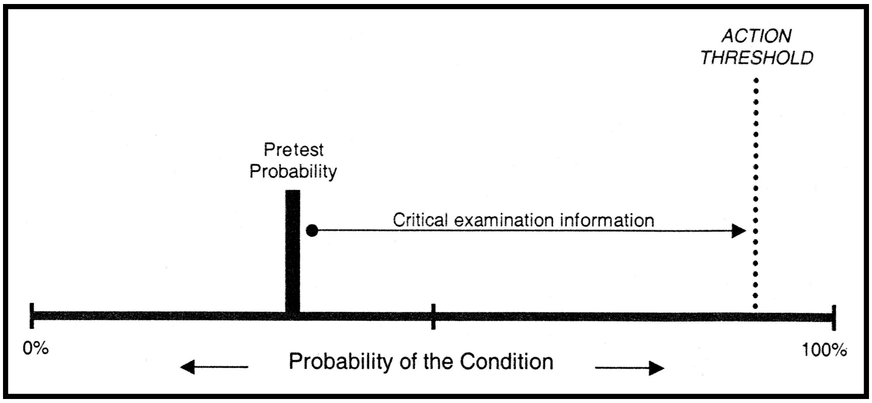

ACTION THRESHOLD
Pretest Probability
Critical examination information
0%
100%
Probability of the Condition

Figure 2.
Illustration of the probabilistic nature to the diagnostic process.[72] The determination of which tests will most efficiently provide the critical examination information is made from the likelihood ratio values.

quantifying the level of certainty in a diagnosis prior to the performance of a test.

The information provided by the results of a diagnostic test will alter the pretest probability to some extent, resulting in a revised posttest probability that the condition of interest is present. The magnitude of the revision is based, as has been noted, on data derived from comparisons of the diagnostic test with a reference standard. Likelihood ratios quantify the direction and magnitude of change in the pretest probability based on the test result and, therefore, provide the best information needed to select the test or tests that will most efficiently move from the uncertainty associated with the pretest probability to the threshold for action.[32,102] To illustrate the process, we will use an example of a question that may arise during the examination of a 67-year-old patient with symptoms in both the low back/buttock and anterior hip/groin that worsen when the patient is walking: Are the patient's symptoms coming from the lumbar spine (eg, lumbar spinal stenosis)?

What is a reasonable pretest probability of lumbar spinal stenosis for this patient? Based on the patient's age and symptoms, epidemiological data[103,104] as well as clinical experience suggest that the probability is fairly high, perhaps 50%. What test should be performed to rule in this diagnosis? Examining the results from several studies[30,105,106] (Tab. 11), the best test appears to be asking the patient whether symptoms are absent when sitting (positive LR=6.6). It is not uncommon that information from the history exceeds that obtained from the systems review or the tests and measurements with regard to diagnostic accuracy. If the test is positive, what should the posttest probability of lumbar spinal stenosis be? Two methods can be used to make this determination. The simpler, but less precise, method uses a nomogram (Fig. 3).[107] A straightedge is anchored along the left-hand side of the nomogram at the point corresponding to the pretest probability. The posttest probability is determined by running the straightedge from this point through the appropriate LR value. The point of intersection of the straightedge with the right-hand side of the nomogram represents the posttest probability.[42]

An alternative method for quantifying posttest probability utilizes a 3-step calculation process described by Sackett et al[42] and outlined below:

1. Convert the pretest probability (50%) to odds using the formula:

$$\textit{Pretest odds}=\frac{\textit{pretest probability}}{1-\textit{pretest probability}}$$

In this example, the pretest odds would be: .50/(1 − .50)=1:1.

2. Multiply the odds by the appropriate LR value (in this case, the positive LR) using the formula:

$$\textit{Pretest odds}\times \textit{LR}=\textit{posttest odds}$$

In this example, the posttest odds would be: 1:1 × 6.6=6.6:1.

3. Convert the posttest odds back to probability using the formula:

$$\frac{\textit{Posttest odds}}{\textit{Posttest odds}+1}=\textit{posttest probability}$$

In this example, the posttest probability would be: 6.6/(6.6 + 1)=87%.

Knowledge of the positive LR values permitted the selection of the test that produced the greatest shift in probability favoring the condition. Had another test been selected with a smaller positive LR, the results would not have been as conclusive. For example, if the therapist had opted to assess pain with lumbar flexion and the test were positive (ie, no pain), the posttest probability would increase to only 58%. Without knowledge of the relative unimportance of this finding, the therapist might over-interpret the value of the positive result.

The importance of the pretest probability is also highlighted by this example. If the patient in question had the same symptoms but was younger, perhaps 45 years of

Table 11.
Accuracy of Diagnostic Tests for Lumbar Spinal Stenosis[a]

Test	Sensitivity	Specificity	Positive LR	Negative LR
Factors from the history				
Symptoms become worse with walking[30]	71 (57, 85)	30 (14, 46)	1.0 (0.80, 1.3)	0.96 (0.50, 1.75)
Ranks standing or walking as worse than sitting with regard to symptoms[105]	89 (76, 100)	33 (12, 55)	1.3 (0.93, 1.9)	0.35 (0.10, 1.2)
Able to walk better when holding on to a shopping cart[105]	63 (42, 85)	67 (40, 93)	1.9 (0.79, 4.5)	0.55 (0.27, 1.1)
Absence of pain when seated[30]	46 (30, 62)	93 (84, 100)	6.6 (2.4, 18.0)	0.58 (0.43, 0.77)
Factors from the examination				
No pain with lumbar flexion[30]	79 (67, 91)	44 (27, 61)	1.4 (1.1, 1.9)	0.48 (0.25, 0.92)
Absent Achilles reflex[30]	46 (31, 61)	78 (64, 92)	2.0 (1.1, 3.6)	0.69 (0.51, 0.95)
Able to walk farther with the spine flexed vs extended[106]	58 (36, 80)	91 (74, 100)	6.4 (0.95, 42.9)	0.46 (0.27, 0.81)

[a] Numbers in parentheses represent 95% confidence intervals calculated from the data presented in the references.

age, the pretest probability of lumbar spinal stenosis would be lower. If the pretest probability was estimated at 20% (pretest odds=0.25:1) and the question of the absence of pain when seated was positive, the posttest probability would increase to 62%. It is likely that, in the mind of the therapist, further confirmation would be needed to reach the action threshold for diagnosing the patient with lumbar spinal stenosis. Based on the data shown in Table 11, comparing walking tolerance with the spine flexed versus extended would be the best option (positive LR=6.4). If this test were positive, the probability would increase from 62% to 91%, likely exceeding the action threshold.

When the pretest probability is low, the therapist may instead seek information to rule out stenosis and then proceed with confirming an alternative hypothesis.[4] In this circumstance, the test with the smallest negative LR would be desirable because a negative result would most effectively exclude the condition. Examining Table 11, it is again apparent that a question from the history will be more effective for this purpose than other factors. The patient is asked to rank sitting, standing, and walking from "best" to "worst" with regard to symptoms. If the test is negative (ie, pain during standing or walking is not ranked as "worst"), the negative LR associated with the finding is 0.33 and the probability of stenosis drops to 8%. Table 11 also illustrates the impact of the phrasing of the question. If the patient is asked simply whether or not symptoms become worse when walking, the result is useless, with positive and negative LR values of about 1.0. If the patient instead is asked about improvement in walking when holding on to a shopping cart, the specificity and positive LR increase, but the negative LR remains fairly low. If the goal is ruling out lumbar spinal stenosis, having the patient rank pain during sitting, standing, and walking has the potential to provide the strongest evidence.

Likelihood ratios provide the most powerful tool for demonstrating the importance of a particular test within the diagnostic process in a quantified manner. Because LR values can be calculated for both positive and negative results, the importance of each can be examined. This is necessary because few tests provide useful information in both capacities, and understanding the relative strength of evidence provided by a negative or positive result helps to refine test interpretation. For these reasons and for other reasons discussed, researchers examining diagnostic tests should calculate, or provide sufficient data to permit the calculation of, LR values.[83] Therapists should focus on LR values in determining which tests are most effective for ruling in or ruling out conditions of interest.

Applying the Evidence—The Consequences of Not Practicing Evidence-Based Diagnosis

Diagnostic tests play a critical role in the management of patients in physical therapy. The results of individual tests are evaluated during the examination process, determining which hypotheses should be ruled in or out, ultimately leading to a decision to a use a certain intervention that is believed to provide optimal outcomes for the patient. The ability to judge evidence for diagnostic tests, select the most appropriate test for an individual patient, and interpret the results will need to become familiar skills if physical therapy diagnosis is to become a more evidence-based process.

Many aspects of physical therapist practice, including diagnosis, have been criticized for excessive allegiance to expert opinion and uncritical acceptance of standards that are not based on evidence.[108,109] Systems of integrating diagnosis and intervention in common usage by physical therapists too frequently owe their popularity to tradition instead of sound data attesting to their usefulness. For example, neurodevelopmental treatment

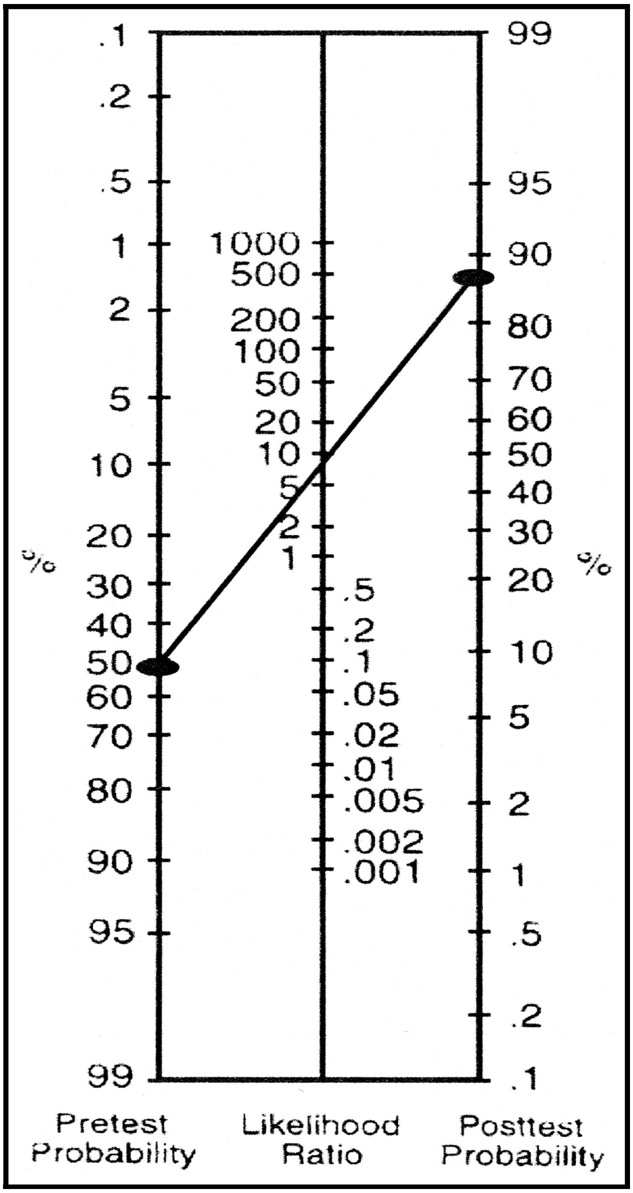


Figure 3.
Nomogram for estimating posttest probability of a diagnosis.[107] For the example given in the test, the pretest probability was estimated as 50%, and the positive likelihood ratio was 6.6.

(NDT) is an approach to the management of patients with movement disorders in which the therapist examines factors such as movement patterns and postural reactions and then selects interventions to reduce abnormal movements and improve function.[110,111] Even though NDT appears to be the method most commonly used by physical therapists for managing children with cerebral palsy,[112] little research has been performed to examine the evidence for examination techniques used within the system or the manner in which the tests are evaluated to determine appropriate interventions.[111]

Without any validation of the diagnostic decision making underlying intervention choices, it is not surprising that clinical trials comparing patients treated with an NDT-based approach versus other interventions have not demonstrated improved outcomes with the use of the NDT system.[113–117] A similar situation exists for the most common treatment approach for patients with LBP, the McKenzie system.[118] The McKenzie system uses a variety of examination techniques, the results of which are used to place patients into categories and to determine interventions. Little work has been done to examine the diagnostic process used by the McKenzie system, and the reliability of the classifications is questionable.[119] A recent clinical trial comparing outcomes for the McKenzie system with chiropractic care and a patient education pamphlet resulted in essentially no differences among the treatment approaches.[120]

Reliance on patient management systems that are not evidence-based, in our view, has negative consequences not only for practitioners, but also for the profession of physical therapy as a whole. Both the McKenzie system and NDT have been used in clinical trials as representative of "physical therapy" interventions for patients with LBP or cerebral palsy, respectively.[117,120] The negative results of these trials have led to the conclusion that physical therapy may not have a role in the management of these conditions. It should not be surprising, however, that systems whose diagnostic procedures are not evidence-based do not result in improved patient outcomes. If diagnostic decisions had been made on the basis of tests with evidence attesting to their ability to focus the examination and determine the most effective interventions, the results might have been more positive. The McKenzie system and NDT serve only to illustrate a more fundamental problem. Without evidence-based diagnosis, interventions will continue to be based on observation that may not even be systematic, pathoanatomical theories, ritual, and opinion. Studies examining the outcomes of such interventions will continue, in our opinion, to offer discouraging results. The solution is not only to explore new and innovative interventions, but to refine the process by which interventions are linked to examination findings by studying evidence-based diagnosis.

Conclusions

The process of diagnosis is an essential task for physical therapists because it serves as the link between examination findings and interventions. To be able to examine diagnosis from an evidence-based perspective, we argue that therapists need to be familiar with the standards defining the "current best evidence" and how the evidence can be used for "making decisions about the care of individual patients."[21(p71)] The standards relate to several aspects of the study design and data analysis.

An important first step is to define the purpose for which a diagnostic test is used. The purpose should be reflected in the choice of a reference standard (measurement) against which the results are compared. Both the diagnostic test and the reference standard should be applied consistently in all subjects and judged by blinded examiners.

The study sample should be representative of the type of patients on whom the test is typically used in the clinical setting. The best statistics for application in individual patient decision making are LRs because they can be used to quantify probability revision based on positive or negative test results. The application of evidence into patient management requires an understanding of probability and the shifts in probability caused by a certain test result. Systems of patient management that link diagnostic tests with interventions may produce less favorable results when the diagnostic process within the system is not evidence-based. More studies are needed to examine commonly used diagnostic methods in physical therapy. The evidence provided by past and future studies should be applied to the management of patients in order to make the practice of physical therapy more evidence-based.

References

1 Rose SJ. Physical therapy diagnosis: role and function. *Phys Ther.* 1989;69:535–537.

2 Sahrmann SA. Diagnosis by the physical therapist: a prerequisite for treatment. *Phys Ther.* 1988;68:1703–1706.

3 Guide to Physical Therapist Practice. 2nd ed. *Phys Ther.* 2001;81:43.

4 Delitto A, Snyder-Mackler L. The diagnostic process: examples in orthopedic physical therapy. *Phys Ther.* 1995;75:203–211.

5 Clinical Research Agenda for Physical Therapy. *Phys Ther.* 2000;80: 499–513.

6 Schwartz JS. Evaluating diagnostic tests: what is done, what needs to be done? *J Gen Intern Med.* 1986;1:266–276.

7 Deyo RA, Haselkorn J, Hoffman R, Kent DL. Designing studies of diagnostic tests for low back pain or radiculopathy. *Spine.* 1994; 19(suppl 18):2057S–2065S.

8 Sackett DL, Wennberg JE. Choosing the best research design for each question: it's time to stop squabbling over the "best" methods. *BMJ.* 1997;315:1636.

9 Shinozaki T, Hasegawa T, Yano E. Ankle-arm index as an indicator of atherosclerosis: it's application as a screening method. *J Clin Epidemiol.* 1998;51:1263–1269.

10 Newman AB, Siscovick DS, Manolio TA, et al. Ankle-arm index as a marker of atherosclerosis in the Cardiovascular Health Study. *Circulation.* 1993;99:837–845.

11 Altman R, Alarcon G, Appelrouth D, et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheum.* 1991;34:505–515.

12 Delitto A, Erhard RE, Bowling RW. A treatment-based classification approach to low back syndrome: identifying and staging patients for conservative management. *Phys Ther.* 1995;75:470–489.

13 Khuffash B, Porter RW. Cross leg pain and trunk list. *Spine.* 1989;14:602–603.

14 Porter RW, Miller CG. Back pain and trunk list. *Spine.* 1986;11: 596–600.

15 McKenzie RA. Manual correction of sciatic scoliosis. *NZ Med J.* 1972;76:194–199.

16 McKenzie RA. *The Lumbar Spine: Mechanical Diagnosis and Therapy.* Waikanae, New Zealand: Spinal Publications Ltd; 1989.

17 Charnley J. Orthopaedic signs in the diagnosis of disc protrusion. *Lancet.* 1951;1:186–192.

18 Deyo RA, Rainville J, Kent DL. What can the history and physical examination tell us about low back pain? *JAMA.* 1992;268:760–765.

19 van den Hoogen HMM, Koes BW, van Eijk JTM, Bouter LM. On the accuracy of history, physical examination, and erthrocyte sedimentation rate in diagnosing low back pain in general practice. *Spine.* 1995;20:318–327.

20 Fritz JM. Use of a classification approach to the treatment of 3 patients with low back syndrome. *Phys Ther.* 1998;78:766–777.

21 Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. *BMJ.* 1996;312:71–72.

22 van Tulder MW, Assendelft WJ, Koes BW, Bouter LM. Method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group for Spinal Disorders. *Spine.* 1997;22:2323–2330.

23 Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ.* 1994;309:1286–1291.

24 Guyatt GH, Sackett DL, Cook DJ. Users' guide to the medical literature, II: how to use an article about therapy or prevention, A: are the results of the study valid? *JAMA.* 1993;270:2598–2601.

25 Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of diagnostic test evaluation. *J Gen Intern Med.* 1989;4:288–295.

26 Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999;282: 1061–1066.

27 Irwig L, Tosteson ANA, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med.* 1994;120:667–676.

28 Begg CB. Methodologic standards for diagnostic test assessment studies. *J Gen Intern Med.* 1988;3:518–520.

29 Armstrong DG, Lavery LA, Vela SA, et al. Choosing a practical screening instrument to identify patients at risk for diabetic foot ulceration. *Arch Intern Med.* 1998;153:289–292.

30 Katz JN, Dalgas M, Stucki G, et al. Degenerative lumbar spinal stenosis: diagnostic value of the history and physical examination. *Arthritis Rheum.* 1995;38:1236–1241.

31 Geddes JR, Harrison PJ. Closing the gap between research and practice. *Br J Psychiatry.* 1997;171:220–225.

32 Jaeschke RZ, Meade MO, Guyatt GH, et al. How to use diagnostic test articles in the intensive care unit: diagnosing weanability using f/Vt. *Crit Care Med.* 1997;25:1514–1521.

33 Task Force on Standards for Measurement in Physical Therapy. Standards for tests and measurements in physical therapy practice. *Phys Ther.* 1991;71:589–622.

34 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? *JAMA.* 1994;271:389–391.

35 Rothstein JM, Echternach JL. *Primer on Measurement: An Introductory Guide to Measurement Issues.* Alexandria, Va: American Physical Therapy Association; 1993:67–73.

36 Ashworth B. Preliminary trial of carisoprodal in multiple scelerosis. *Practitioner.* 1964;192:540–542.

37 Pandyan AD, Johnson GR, Price CI, et al. A review of the properties and limitations of the Ashworth and modified Ashworth Scales as measures of spasticity. *Clin Rehabil.* 1999;13:373–383.

38 Haas BM, Bergstrom E, Jamous A, Bennie A. The inter-rater reliability of the original and of the modified Ashworth scale for the assessment of spasticity in patients with spinal cord injury. *Spinal Cord.* 1996;34:560–564.

39 Allison SC, Abraham LD, Petersen CL. Reliability of the Modified Ashworth Scale in the assessment of plantarflexor muscle spasticity in patients with traumatic brain injury. *Int J Rehabil Res.* 1996;19:67–78.

40 Katz RT, Rovai GP, Bait C, Rymer WZ. Objective quantification of spastic hypertonia: correlation with clinical findings. *Arch Phys Med Rehabil.* 1992;73:339–347.

41 Fowler EG, Nwigwe AI, Ho TW. Sensitivity of the pendulum test for assessing spasticity in persons with cerebral palsy. *Dev Med Child Neurol.* 2000;42:182–189.

42 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine.* 2nd ed. Boston, Mass: Little, Brown and Co Inc; 1992.

43 Burke DT, Burke MA, Bell R, et al. Subjective swelling: a new sign for carpal tunnel syndrome. *Am J Phys Med Rehabil.* 1999;78:504–508.

44 Cibulka MT, Koldehoff R. Clinical usefulness of a cluster of sacroiliac joint tests in patients with and without low back pain. *J Orthop Sports Phys Ther.* 1999;29:83–92.

45 Levangie PK. Four clinical tests of sacroiliac joint dysfunction: the association of test results with innominate torsion among patients with and without low back pain. *Phys Ther.* 1999;79:1043–1057.

46 Dreyfuss P, Michaelsen M, Pauza K, et al. The value of medical history and physical examination in diagnosing sacroiliac joint pain. *Spine.* 1996;21:2594–2602.

47 Maigne J-Y, Aivaliklis A, Pfefer F. Results of sacroiliac joint double block and value of sacroiliac pain provocation tests in 54 patients with low back pain. *Spine.* 1996;21:1889–1892.

48 Slipman CW, Sterenfeld EB, Chou LH, et al. The predictive value of provocative sacroiliac joint stress maneuvers in the diagnosis of sacroiliac joint syndrome. *Arch Phys Med Rehabil.* 1998;79:288–292.

49 Cibulka MT, Delitto A, Koldehoff RM. Changes in innominate tilt after manipulation of the sacroiliac joint in patients with low back pain: an experimental study. *Phys Ther.* 1988;68:1359–1363.

50 Panzer RJ, Suchman AL, Griner PF. Workup bias in prediction research. *Med Decis Making.* 1987;7:115–119.

51 Rubenstein RA, Shelbourne KD, McCarroll JR, et al. The accuracy of the clinical examination in the setting of posterior cruciate ligament injuries. *Am J Sports Med.* 1994;22:550–557.

52 Pinto-Martin JA, Torre C, Zhao H. Nurse screening of low-birth-weight infants for cerebral palsy using goniometry. *Nurs Res.* 1997;46:284–287.

53 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA.* 1995;274:645–651.

54 Ransohoff DF, Feinstein AR. Problems of spectrum bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299:926–930.

55 Itoh H, Kurosaka M, Yoshiya S, et al. Evaluation of functional deficits determined by four different hop tests in patients with anterior cruciate ligament deficiency. *Knee Surg Sports Traumatol Arthrosc.* 1998;6:241–245.

56 Greenhalgh T. How to read a paper: papers that report diagnostic or screening tests. *BMJ.* 1997;315:540–543.

57 Ganko A, Engebretsen L, Ozer H. The rolimeter: a new arthrometer compared with the KT-1000. *Knee Surg Sports Traumatol Arthrosc.* 2000;8:36–39.

58 Liu SH, Osti L, Henry M, Bocchi L. The diagnosis of acute complete tears of the anterior cruciate ligament: comparison of MRI, arthrometry and clinical examination. *J Bone Joint Surg Br.* 1995;77:586–588.

59 Snyder-Mackler L, Fitzgerald GK, Bartolozzi AR, Ciccotti MG. The relationship between passive joint laxity and functional outcome after anterior cruciate ligament injury. *Am J Sports Med.* 1997;25:191–195.

60 Katz JW, Fingeroth RJ. The diagnostic accuracy of ruptures of the anterior cruciate ligament comparing the Lachman test, the anterior drawer sign, and pivot shift test in acute and chronic knee injuries. *Am J Sports Med.* 1986;14:88–91.

61 Cooperman JM, Riddle DL, Rothstein JM. Reliability and validity of judgments of the integrity of the anterior cruciate ligament of the knee using the Lachman's test. *Phys Ther.* 1990;70:225–233.

62 Long AL. The centralization phenomenon: its usefulness as a predictor of outcome in conservative treatment of chronic low back pain (a pilot study). *Spine.* 1995;20:2513–2521.

63 Karas R, McIntosh G, Hall H, et al. The relationship between nonorganic signs and centralization of symptoms in the prediction of return to work for patients with low back pain. *Phys Ther.* 1997;77:354–360.

64 Werneke M, Hart DL, Cook D. A descriptive study of the centralization phenomenon: a prospective analysis. *Spine.* 1999;24:676–683.

65 Fritz JM, Delitto A, Vignovic M, Busse RG. Inter-rater reliability of judgments of the centralization phenomenon and status change during movement testing in patients with low back pain. *Arch Phys Med Rehabil.* 2000;81:57–61.

66 Sheps SB, Schechter MT. The assessment of diagnostic tests: a survey of current medical research. *JAMA.* 1984;252:2418–2422.

67 Egglin TK, Feinstein AR. Context bias: a problem in diagnostic radiology. *JAMA.* 1996;276:1752–1755.

68 Begg CB. Biases in the assessment of diagnostic tests. *Stat Med.* 1987;6:411–423.

69 Gellman H, Gelberman RH, Tan AM, Botte MJ. Carpal tunnel syndrome: an evaluation of the provocative diagnostic tests. *J Bone Joint Surg Am.* 1986;68:735–737.

70 Kuhlman KA, Hennessey WJ. Sensitivity and specificity of carpal tunnel syndrome signs. *Am J Phys Med Rehabil.* 1997;76:451–457.

71 Kassirer JP. Our stubborn quest for diagnostic certainty: a cause of excessive testing. *N Engl J Med.* 1989;320:1489–1491.

72 Bernstein J. Decision analysis. *J Bone Joint Surg Am.* 1997;79:1404–1414.

73 Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures: principles and applications. *Ann Intern Med.* 1981;94:557–592.

74 Hagen MD. Test characteristics: how good is that test? *Med Decis Making.* 1995;22:213–233.

75 Lauder TD, Dillingham TR, Andary M, et al. Effect of history and exam in predicting electrodiagnostic outcome among patients with suspected lumbosacral radiculopathy. *Am J Phys Med Rehabil.* 2000;79:60–68.

76 Kortelainen P, Puranen J, Koivisto E, Lähde S. Symptoms and signs of sciatica and their relation to the localization of the lumbar disc herniation. *Spine.* 1985;10:88–92.

77 Sackett DL. A primer on the precision and accuracy of the clinical examination. *JAMA.* 1992;267:2638–2644.

78 Schulzer M. Diagnostic tests: a statistical review. *Muscle Nerve.* 1994;17:815–819.

79 Calis M, Akgün K, Birante M, et al. Diagnostic value of clinical diagnostic tests in subacromial impingement syndrome. *Ann Rheum Dis.* 2000;59:44–47.

80 Dujardin B, Van den Ende J, Van Gompel A, et al. Likelihood ratios: a real improvement for clinical decision making? *Eur J Epidemiol.* 1994;10:29–36.

81 Lurie JD, Sox HC. Principles of medical decision making: spine update. *Spine.* 1999;24:493–498.

82 Boyko EJ. Ruling out or ruling in disease with the most sensitive or specific diagnostic test. *Med Decis Making.* 1994;14:175–179.

83 Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: What are the results and will they help me in caring for my patients? *JAMA.* 1994;271:703–707.

84 Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test results. *J Clin Epidemiol.* 1991;44:763–770.

85 Riddle DL, Stratford PW. Interpreting validity indexes for diagnostic tests: an illustration using the Berg Balance Test. *Phys Ther.* 1999;79:939–948.

86 Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics With Confidence.* 2nd ed. London, England: BMJ Books; 2000.

87 Sackett DL, Strauss SE, Richardson WS, et al. *Evidence-based Medicine. How to Practice and Teach EBM.* 2nd ed. Edinburgh, Scotland: Churchill Livingstone; 2000:233–243.

88 Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med.* 1998;17:857–872.

89 Harper R, Reeves B. Reporting of precision of estimates for diagnostic accuracy: a review. *BMJ.* 1999;318:1322–1333.

90 Glass GV, Hopkins KD. *Statistical Methods in Education and Psychology.* 3rd ed. Boston, Mass: Allyn & Bacon; 1995.

91 Goodman SN. Toward evidence-based medical statistics, 1: the P-value fallacy. *Ann Intern Med.* 1999;130:995–1004.

92 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.

93 Grove WM, Andreasen NC, McDonald-Scott P, et al. Reliability studies of psychiatric diagnosis: theory and practice. *Arch Gen Psychiat.* 1981;38:408–413.

94 Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. *Arch Gen Psychiat.* 1985;42:725–728.

95 Smieja M, Hunt DL, Edelman D, et al. Clinical examination for the detection of protective sensation in the feet of diabetic patients. *J Gen Intern Med.* 1999;14:418–424.

96 Sox HC Jr. Probability theory in the use of diagnostic tests: an introduction to critical study of the literature. *Ann Intern Med.* 1986; 104:60–66.

97 Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med.* 1980;302:1109–1117.

98 Pauker SG, Kassirer JP. Therapeutic decision-making: a cost benefit analysis. *N Engl J Med.* 1975;293:229–234.

99 McNeil BJ, Pauker SG. The patient's role in assessing the value of diagnostic tests. *Radiology.* 1979;132:605–610.

100 van der Lee JH, Wagenaar RC, Lankhorst GJ. Forced use of the upper extremity in chronic stroke patients: results from a single-blind randomized clinical trial. *Stroke.* 1999;30:2369–2375.

101 Ross JM, Sox HC. If at first you don't succeed: clinical problem-solving. *N Engl J Med.* 1995;333:1557–1560.

102 Hayden SR, Brown MD. Likelihood ratio: a powerful tool for incorporating the results of a diagnostic test into clinical decision-making. *Ann Emerg Med.* 1999;33:575–580.

103 Fritz JM, Delitto A, Welch WC, Erhard RE. Lumbar spinal stenosis: a review of current concepts in evaluation, management, and outcome measurements. *Arch Phys Med Rehabil.* 1998;79:700–708.

104 Herno A, Partanen K, Talaslahti T, et al. Long-term clinical and magnetic resonance imaging follow-up assessment of patients with lumbar spinal stenosis after laminectomy. *Spine.* 1999;24:1533–1537.

105 Fritz JM, Erhard RE, Delitto A, et al. Preliminary results of the use of a two-stage treadmill test as a clinical diagnostic tool in the differential diagnosis of lumbar spinal stenosis. *J Spinal Dis.* 1997;10:410–416.

106 Dong G, Porter RW. Walking and cycling tests in neurogenic claudication. *Spine.* 1989;14:965–969.

107 Fagan TJ. Nomogram for Bayes's theorem. *N Engl J Med.* 1975; 293:257.

108 Rothstein JM. Editor's note: questions for the disciples. *Phys Ther.* 1994;74:694–696.

109 Fritz JM, Delitto A, Erhard RE, Roman M. An examination of the selective tissue tension scheme, with evidence for the concept of a capsular pattern of the knee. *Phys Ther.* 1998;78:1046–1061.

110 Keshner EA. Reevaluating the theoretical method underlying the neurodevelopmental theory: a literature review. *Phys Ther.* 1981;61: 1035–1040.

111 DeGangi GA, Royeen CB. Current practice among neurodevelopmental treatment association members. *Am J Occup Ther.* 1994;48: 803–808.

112 Bly L. A historical and current view of the basis of NDT. *Pediatric Physical Therapy.* 1991;3:131–135.

113 Fetters L, Kluzik J. The effects of neurodevelopmental treatment versus practice on the reaching of children with spastic cerebral palsy. *Phys Ther.* 1996;76:346–358.

114 Wagenaar RC, Meijer OC, van Wieringen PC. The functional recovery of stroke: a comparison between neuro-developmental treatment and the Brunnstrom method. *Scand J Rehabil Med.* 1990;22:1–8.

115 Law M, Russell D, Pollock N, et al. A comparison of intensive neurodevelopmental therapy plus casting and a regular occupational therapy program for children with cerebral palsy. *Dev Med Child Neurol.* 1997;39:664–670.

116 Law M, Cadman D, Rosenbaum P, et al. Neurodevelopmental therapy and upper-extremity inhibitive casting for children with cerebral palsy. *Dev Med Child Neurol.* 1991;33:379–387.

117 Palmer FB, Shapiro BK, Watchel RC, et al. The effects of physical therapy on cerebral palsy: a controlled trial in infants with spastic diplegia. *N Engl J Med.* 1988;318:803–808.

118 Battié MC, Cherkin DC, Dunn R, et al. Managing low back pain: attitudes and treatment preferences of physical therapists. *Phys Ther.* 1994;74:219–226.

119 Riddle DL, Rothstein JM. Intertester reliability of McKenzie's classifications of the syndrome types present in patients with low back pain. *Spine.* 1993;18:1333–1344.

120 Cherkin DC, Deyo RA, Battié M, et al. A comparison of physical therapy, chiropractic manipulation, and provision of an educational booklet for the treatment of patients with low back pain. *N Engl J Med.* 1998;339:1021–1029.