Physical Medicine and Rehabilitation - International

Austin Publishing Group

Open Access Full Text Article

Research Article

Diagnostic Test Principles: Overview and Applications for Pain Medicine

Upadhye S[1] and Kumbhare DA[2]*

[1]Department of Emergency Medicine, McMaster University, Canada

[2]Department of Medicine, University of Toronto, Canada

***Corresponding author:** Kumbhare DA, Department of Medicine, University of Toronto, Canada

Received: July 03, 2014; **Accepted:** July 20, 2014; **Published:** Aug 01, 2014

Abstract

Patients frequently present with undiagnosed medical complaints. Physician's assessments involve a history, physical examination and review of relevant investigations before arriving at a diagnosis. We present a clinical scenario of lower back pain and assist in the derivation of the diagnosis based upon information available in the literature. Common clinical issues are discussed from a statistical decision perspective with aids for the busy clinician. The need for diagnostic tests in decision making is discussed with an outline of key concepts such as measures of test efficacy, recognizing biases in diagnostic studies, critical appraisal tools. Common clinical situations such as ordering a test for patient reassurance, over-reliance on a diagnostic test are also discussed.

Introduction

Patients frequently present to their physician with undiagnosed medical complaints, and it is up to the physician to clinically assess these patients and undertake treatment plans. Part of this process includes doing a proper history, physical examination, and ordering further diagnostic tests as warranted. Diagnostic tests can be useful to help confirm or refute a medical diagnosis (in conjunction with clinical judgment), determine the severity of disease, and/or evaluate responses to treatment once a diagnosis has been made [1]. It is imperative to understand that diagnostic testing should not replace clinical judgment, as the test results may not be infallible, and if interpreted incorrectly, can be misleading. A key tenet of diagnostic testing is that tests should be ordered ONLY when there is a potential change in management decision-making, not just for the sake of confirming that which the physician already knows or is planning to do, or to put the patient at ease.

Case Scenario

You have been referred a patient with non-traumatic Low Back Pain (LBP) for further assessment and management from a community nurse practitioner. You complete a thorough history and physical examination, and determine that there are no "red flags" for low back pain that merit emergent hospital referral [2]. On further questioning, there is reluctance to go back to work and there are some obvious "yellow flags" for long-term disability, chronicity or work loss [3]. Finally, there are no objective findings of neurologic deficit that merit immediate advanced imaging.

You and your patient agree on an initial conservative management plan, but the patient would like confirmatory diagnostic imaging just to "see what's going on." You explain that there is no role for X-rays or CT scan due to their lack of utility in non-traumatic LBP (and potential harms). The patient then pushes you to order an MRI. You know this is not warranted at this time, as this is contrary to current LBP guidelines [2], and have been over utilized for LBP in Canada [4]. Furthermore, you don't feel an MRI will change your initial management plan, and the patient may be fishing for a diagnosis to justify a disability claim submission. You discuss this with your patient.

Key Reference

You are concerned that your otherwise healthy patient may be looking to go on disability and not return to work. You are aware of a recent review on this concern regarding predictors of persistent chronic disabling low back pain [3]. The key results from this systematic review are summarized in Table 1.

The Need for Diagnostic Tests – Key Concepts Outline

Decision threshold lines

As previously stated, clinicians use (or should use) diagnostic tests to help them make decisions (or share decision-making) with patients that alter management. There are occasions when a clinical diagnosis is so obvious that a diagnostic test is not warranted and the clinician has the information necessary to precede directly to management. For example, a patient who comes into your office with an obviously dislocated finger joint does not need an X-ray to confirm dislocation; they need a local anesthetic, relocation and splinting. An X-ray may be warranted to determine if there is a concurrent fracture that may require an operative fixation after the finger has been relocated but this is a different diagnostic and management issue. In this situation, the diagnostic test (finger X-ray) is ordered to make a decision of potential operative fixation, not to confirm a clinical dislocation and confirmation of relocation after the finger was clinically reduced.

Every clinical assessment by a clinician about a specific patient problem should be an exercise in diagnostic decision-making. Each question asked in the history and each physical examination maneuver is used to sequentially to increase or decrease: (a) the likelihood of a diagnostic disease probability, the thresholds for which a decision can be made to discard the disease diagnosis/treatment plan (and pursue another diagnostic possibility), or (b) confirm the Disease of Interest (DoI) and start a management plan. A continuum of disease pretest probability can be visualized as a clinical decision line, anchored on

Citation: Upadhye S and Kumbhare DA. Diagnostic Test Principles: Overview and Applications for Pain Medicine. Phys Med Rehabil Int. 2014;1(1): 8.

Table 1: Predictors of persistent disabling low back pain [3]. Health status Predictors.

Definition	No. of studies	Median Positive LR (Range)	Median Negative LR (Range)
Health Status Predictors			
Lower vs. better health status (3- 6mo)	3	1.6 (1.1-1.7)	0.73 (0.66-0.86)
Lower vs. better health status (1 year)	5	1.8 (1.1-2.0)	0.85 (0.56-0.99)
Higher vs. lower psychiatric comorbidity scores (3-6mo)	4	1.9 (1.4-2.1)	0.69 (0.55-0.85)
Higher vs. lower psychiatric comorbidity scores (1 yr)	4	2.2 (1.9-2.3)	0.85 (0.55-0.93)
Prior LBP episodes: more vs. less/no episodes (3-6mo)	6	1.0 (0.90-1.2)	0.88 (0.53-1.1)
Prior LBP episodes: more vs. less/no episodes (1 yr))	5	1.1 (0.95-1.2)	0.81 (0.21-1.1)
Clinical Signs and Symptoms			
Radiculopathy/leg pain vs. no leg pain/radiculopathy (3-6mo)	5	1.4 (1.1-1.7)	0.63 (0.52-0.93)
Radiculopathy/leg pain vs. no leg pain/radiculopathy (1 yr)	7	1.4 (1.2-2.4)	0.82 (0.54-0.94)
Nonorganic signs/somatization vs. none (3mo)	1	2.5 (95% CI 1.8-3.4)	0.8 (95%CI 0.74-0.89)
Widespread pain/somatization vs none (1 yr)	3	3.0 (1.7-4.6)	0.71 (0.31-0.76)
Pain percpetions/avoidance		Median LR (Range)	
Baseline pain intensity (3-6mo)	6	High: 1.7 (1.1-3.7) Medium: 0.86 (0.66-1.2) Low: 0.70 (0.07-0.86)	
Baseline pain intensity (1 yr)	3	High: 1.3 (1.2-2.0) Medium: 0.78 (0.72-1.2) Low: 0.33 (0.08-0.97)	
Baseline function impairment (3- 6mo)	6	High: 1.4 (1.3-3.5) Medium:1.3 (0.74-1.5) Low: 0.53 (0.18-1.1)	
Baseline function impairment (1 yr)	3	High: 2.1 (1.2-2.7) Medium: 0.86 (0.85-1.7) Low: 0.40 (0.10-0.52)	
Fear avoidance behavior intensity (3-6mo)	4	High: 2.2 (1.5-4.9) Medium: 1.1 (1.0-1.5) Low: 0.46 (0.30-0.73)	

the left with a 0% probability of disease, and the right with a 100% probability (Figure 1). A "negative" decision threshold (T_{NEG}) is that point of pretest probability on a clinical decision line, below which the probability of disease of interest is so low that the diagnosis can be discarded, and no testing is required to confirm or refute that diagnosis. Similarly, the "positive" decision threshold (T_{POS}) is that point on the decision line that the pretest probability of DoI is so likely that a diagnostic test is not needed, and a treatment plan based on that DoI can be initiated. Predictably, there is a zone of uncertainty between these two thresholds that, after a clinical assessment has failed to cross a threshold, a diagnostic test (or sequential tests) may be required to move past one of the decision thresholds.

Understandably, depending on the DoI under consideration, the decision thresholds may move to the left and right. For example, a physician doing a clinical assessment of a patient with potential fibromyalgia may not worry about whether or not a screening tool actually diagnoses the disease before making a decision to treat (lower T_{POS}) of not treat (higher T_{NEG}), since the consequences of diagnostic error may not be necessarily high. On the other hand, if the physician has a patient with persistent headaches is worried about a brain tumor, they may feel that a clinical examination is not good enough to rule out the diagnosis, and require an advanced imaging test to rule this out (i.e. Very low T_{NEG}, need a scan to cross below this threshold to rule out). The same may apply in a situation where a diagnostic test is required to cross T_{POS} and start treating the DoI.

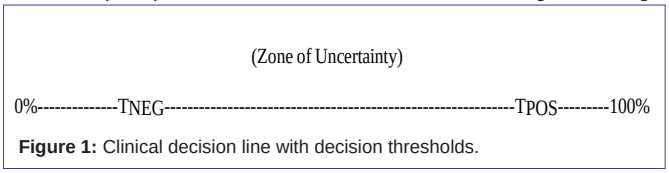

Figure 1: Clinical decision line with decision thresholds.

Measures of test efficacy

An important issue in understanding diagnostic tests is the notion that any measurable parameter in a general population will be either "normal" or "abnormal." Rarely is a measurement outcome in medicine dichotomous (i.e. dead vs. alive); rather, many tests involve a range of continuous values, with defined cutoffs signifying whether the result is "normal" or "abnormal" (e.g. blood pressure, Hb level, or height). A "normal" range of test values encompasses the "average" value for a disease-free population, and the range of 2 standard deviations (95%) on either side of this average for a normal population distribution. Note that the remaining 5% of patients lying outside of this range may have an "abnormal" test value but are still disease-free. Similarly, a patient may have a test result which may be within this 95% "normal" range, yet have evidence of disease (especially if that test result is significantly different from previous values for the same individual). This is illustrated in Figure 2.

Generally, the performance characteristics of a diagnostic test are reported in comparison to a "gold" standard for the disease diagnosis. There are few absolute gold standards, however, in clinical medicine; for example, an ultrasound may be a useful test to diagnose acute appendicitis, but the gold standard is the operative and pathologic diagnosis. In most cases, physicians are left with a clinical outcome

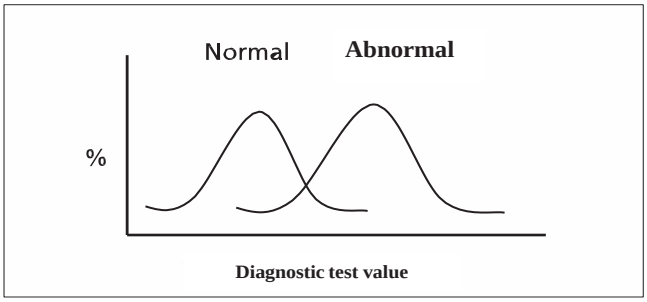

Figure 2: The distribution of normal and abnormal diagnostic test values in patients, with some overlap and variability between the disease free population and the disease population. Some values can fall in either health or disease.

over time, and must rely on the diagnostic test characteristics with this outcome in mind. As such, the test results will be reported as true positive, true negative, false positive or false negative. This can only occur after the final clinical outcome has been determined. These parameters are summarized in the typical 2x2 table for comparing test performance (Table 2).

The standard 2x2 table is organized based on what the test results say about presence or absence of disease, compared to the actual truth based on gold standard testing comparison. There will be true results (true positive [A] and true negative [D]) and false results (false positive [B] and false negative [C]). The accuracy of a test is the proportion of true results amongst all results (i.e. [A+D]/[A+B+C+D]). This has little clinical meaning; however, as clinicians ordering tests are not generally interested in how accurate a test is. They are generally more interested in a test's ability to rule in or rule out a disease state. The more practical determinants of test utility are sensitivity and specificity.

A test with high specificity is one that, if positive, will rule "in" a disease. This is represented by the mnemonic "SpIn" (**S**pecific test, positive result, disease ruled in). In the 2x2 table, this is represented by a high value for cell D, which is the true negative rate. The higher the true negative rate D, the higher the overall specificity will be [D/(B+D)], and a positive test will likely therefore be truly positive. A test with high sensitivity, on the other hand, has the ability to rule "out" a disease if the test is negative. This is remembered with the mnemonic "SnOut" (Sensitive test, negative result, disease ruled OUT). In the table, this is represented by a high value of cell A, the true positive rate. The higher the value of A, the overall sensitivity is higher [A/(A+C)], and a negative test is more likely truly negative.

Table 2: Diagnostic test parameters.

	Patient Disease status (actual)				
Result of Diagnostic Test	Test Positive	Disease present	Disease absent	Totals	
	Test Negative	True Positives (A)	False Positives (B)	With positive tests (A+B)	PPV = A/(A+B)
	Totals	False Negatives (C)	True Negatives (D)	With negative tests (C+D)	NPV = D/(C+D)
		With Disease (A+C)	Without Disease (B+D)	A+B+C+D	
		Sensitivity = A/(A+C)	Specificity = D/(B+D)		

Accuracy = (A+D) / (A+B+C+D)

The challenge with many diagnostic tests, of course, is that they do not necessarily generate binary outcomes. As previously mentioned, many clinical items are measured on a continuous scale (e.g. Blood pressure, Hb level, etc.), and the difference between normal/abnormal (or disease-free/disease-present) is set by cutoff thresholds ideally based on clinical observations (but often arbitrarily). Changing the cutoff of normal/abnormal measures will change the numbers occupying the cells of the 2x2 table describing the test characteristics in relation to the disease presence. There is a trade-off in sensitivity and specificity as a result of raising or lowering numeric cutoff thresholds. Lowering a numeric cutoff threshold will result in more "positive" tests, but not necessarily more disease cases, which will therefore increase sensitivity but reduce specificity. Contrarily, increasing a numeric cutoff threshold will increase the "negative" test rate, but not necessarily less disease cases, which will increase specificity but lower sensitivity. The relationship between sensitivity and specificity at different cutoff thresholds can be described graphically with a Receiver Operator Characteristic (ROC) curve [5]. The "Area under Curve" (AUC) of a ROC is maximized when the cutoff is set at that threshold occupying the upper leftmost point on the curve, where sensitivity and specificity have been mutually maximized. When the test has both 100% sensitivity and specificity at an optimal cutoff, the corresponding AUC will be 1.0 (perfect discriminative value); such an ideal test almost never exists. A straight diagonal line from bottom left to top right will have an AUC of 0.5, and therefore no discriminative value (Sense 50% and Spec 50%; Figure 3). In general, the higher the AUC, the more discriminative the test; most authors advise that a minimum acceptable AUC of ≥0.75 should be acceptable for most clinicians. Keep in mind that a ROC curve may have one set of characteristics if the test has been applied under a single set of conditions, but may not perform the same when applied under slightly varied conditions. Clinicians may find a summary ROC curve generated by pooling average sensitivity and specificity from a variety of diagnostic studies more reliable and trustworthy for clinical decision-making, rather than a ROC idealized in a single study that may not be generalizable to your patient population. Finally, the acceptability of a test cutoff threshold, the corresponding sensitivity and specificity values, and ROC AUC will depend upon the clinical

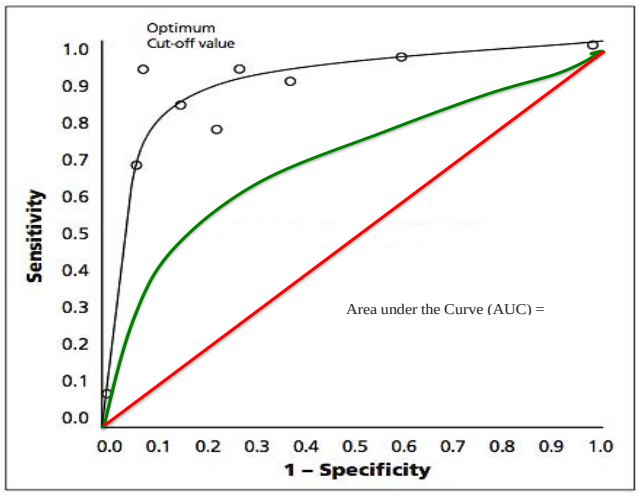

Figure 3: Receiver operator characteristic curves.

context of your specific patient. For example, an AUC for straight leg raise of 0.76 may be acceptable to confirm or refute a clinical diagnosis of lumbar disc hernia ion in a low back pain patient, but another physical examination test with the same AUC of 0.76 may not be acceptably discriminating enough when considering a diagnosis of spinal epidural abscess in the same low back pain patient. The latter scenario would certainly require a test of much higher discriminatory power, given the importance of making (or not missing!!) this diagnosis.

As seen in Table 2, the columns of the table will define sensitivity and specificity. The rows, on the other hand, will define positive and Negative Predictive Values (PPV and NPV). These are commonly reported test characteristics that unfortunately have the potential to be misleading. While predictive values may help to provide a probability of disease state based on test result (PPV = Probability of disease present when test positive, NPV = probability of disease absent when test negative), these values may be biased based on pretest probability (i.e. prevalence) of the disease state before the test is done. While sensitivity and specificity are relatively stable metrics with respect to prevalence, PPV and NPV can change significantly [1,6]. This is illustrated in Tables 3a and 3b. Table 3a represents a hypothetical population of back pain patients seen in a general practice office, who are being evaluated by physical examination for possible spinal stenos, is. Table 3b represents a hypothetical referral population also being evaluated with physical examination in a specialist spine clinic for possible spinal stenos is. As can be seen, the proportions of patients in various cells of the 2x2 are quite different in both populations, and the prevalence is quite different in both groups. However, even though sensitivity and specificity are the same in both groups, one can see that PPV and NPV are very different. Depending on where a clinician works and what the prevalence (pretest probability) of the disease of interest is PPV and NPV may be useful or not useful. As can be seen, in populations with low prevalence of disease (Table 3a), the diagnostic test may suffer poor PPV yet very good NPV. Conversely, in a high prevalence population (Table 3b), the opposite is true (i.e. Good PPV, poor NPV). In situations where the disease prevalence may be uncertain (likely most of the time for many clinicians), it is hard to put the PPV and NPV characteristics in context. In such cases, it is advisable to stick with sensitivity and specificity in isolation. If there is a reasonable estimate of prevalence, however, and the clinician wants to avoid using prevalence-biased PPV and NPV values, the best diagnostic test metric is the likelihood ratio.

Tables 3a & 3b: Influence of prevalence on diagnostic test parameters.
Table 3a: General practice physician's office.

	Disease present (SS)	Disease absent (SS)	Totals
Test positive (PE)	3	47	50
Test negative (PE)	3	47	50
Totals	6	94	100

Sensitivity = A/ (A+C) = 3/(3+3) = 50% Specificity = D/(B+D) = 47/(47+47) = 50%
PPV = A/(A+B) = 3/(3+47) = 6%
NPV = D/(C+D) = 47/ (3+47) = 94%
Prevalence of disease = 6%
PE = Physical Examination, SS = Spinal Steno sis

Table 3b: Spine specialist referral clinic.

	Disease present (SS)	Disease absent (SS)	Totals
Test positive (PE)	45	5	50
Test negative (PE)	45	5	50
Totals	90	10	100

Sensitivity = A/ (A+C) = 45/(45+45) = 50% Specificity = D/(B+D) = 5/(5+5) = 50%
PPV = A/(A+B) = 45/(45+5) = 90%
NPV = D/(C+D) = 5/ (45+5) = 10%
Prevalence of disease = 90%
PE = Physical Examination, SS = Spinal Steno sis

Likelihood Ratios (LRs) are the best indicators of a diagnostic test's utility in the context of pretest probability, in order to facilitate clinical decision-making (after determining post-test probability). A positive LR (LR+) is the ratio of true-positive rate to false-positive rate, and the negative LR (LR-) is the ratio of false-negative rate to true-negative rate. As such, they are simple calculations based on sensitivity and specificity as follows:

Positive LR (LR+) = Sensitivity/ (1-Specificity)

Negative LR (LR-) = (1-Sensitivity)/Specificity

In general, as LR+ increases, the test becomes a stronger positive predictor and, in reverse, the lower the LR- value, the stronger negative prediction by the test. Ideal LR+ values should be >10, with a strong and decisive change in post-test probability (i.e. enough to move beyond T_{POS} on the decision threshold line and start management of confirmed disease. LR>20 are considered definitively ruled in). Similarly, to rule out a disease (i.e. move below T_{NEG} threshold), LR- values should be <0.1 (ideally <0.05 to definitively rule out disease). Understanding the formulas above, a highly specific test with a subsequently high LR+ would be useful to rule in a disease if positive (SpIn), whereas a highly sensitive test with subsequently low LR- should be useful to rule out a disease if negative (SnOut).

Likelihood ratios can be used in conjunction with pretest probability (PreTP) to generate a posttest probability (PostTP). This is classically achieved using a Fagan nomogram (Figure 4). Four scenarios are presented in the example. In scenario 1), a low PreTP disease patient (abdominal aneurysm) is ordered an excellent discriminatory test (CT scan; LR+ 10, LR- 0.05) which is negative, resulting in a very low PostTP of 0.5%; this is likely enough to rule out aneurysm in this patient (i.e. move below T_{NEG}). In scenario 2), a high PreTP patient (disc herniation) is ordered an excellent discriminatory test (MRI; LR+ 10, LR- 0.5), test is positive, resulting in a high PostTP of 90%; disc herniation is confirmed (move beyond T_{POS}). In scenario 3), a patient with high PreTP DVT gets an excellent discriminatory test (LR+ 18, LR- 0.5) but test is negative, resulting in a still high PostTP 20%; DVT is still a diagnostic consideration and another test may be required before a clinical decision can be made. In scenario 4), a patient with low PreTP (10%) of appendicitis is ordered a poorly discriminatory test (WBC LR+ 2.2, LR- 0.18) and test is positive, resulting in a PostTP of approximately 20%; appendicitis is now higher probability compared to PreTP, and further testing is warranted to refute or confirm the diagnosis. The first 2 scenarios highlight the benefits of using highly discriminatory tests and how to manage appropriate outcomes. Scenario 3) highlights a discrepancy when a high PreTP condition is matched to a properly discriminatory test, yet the result is discordant to PreTP. In this case, the clinician may have to reevaluate the PreTP determination, reassess the test to ensure accurate interpretation, or order another test to confirm the initial evaluation. Scenario 4) highlights the problems of ordering

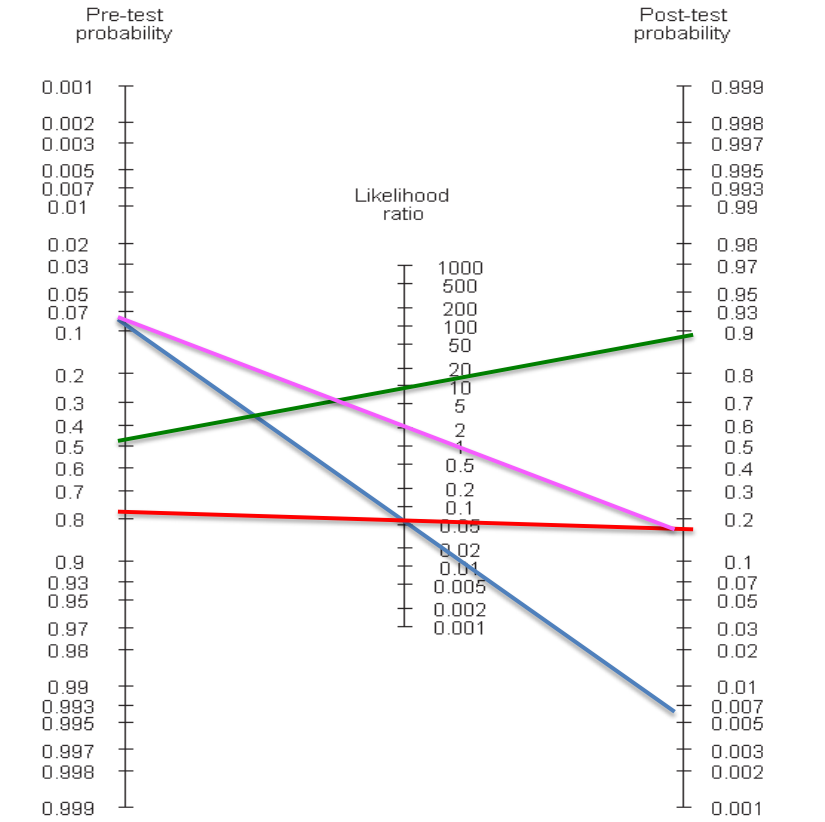

Figure 4: Fagan nomogram use with likelihood ratios

Scenario 1: Patient with abdominal pain has a low PreTP of abdominal aneurysm (10%), gets a CT scan of abdomen (LR+ 10, LR- 0.05), test is negative; the Post TP of aneurysm is only 0.5% (blue line). Aneurysm may be safely ruled out at this point.

Scenario 2: Patient with suspected disc herniation for low back pain (PreTP 50%), gets an MRI of spine (LR+ 10, LR- 0.5), test is positive; the PostTP of disc herniation is 90% (green line). Disc herniation is likely confirmed after this test.

Scenario 3: Patient with suspected clinical DVT has high PreTP (80%), gets an US of calf (LR+ 18, LR- 0.5), test is negative; the PostTP of DVT is still almost 20% (red line). DVT is probably not yet ruled out.

Scenario 4: Patient with suspected appendicitis has PreTP of 10%, and has a white blood cell count done (LR+ 2.2, LR- 0.18), test is positive; PostTP is now almost 20% for appendicitis (pink line). This not likely enough to move past either decision threshold will need another (more discriminative) test.

poorly discriminating tests in low PreTP situations. A test with weak LR+ or LR- characteristics are generally useless and not worth ordering, as they inevitably don't lead to a PostTP that allows crossing a decision threshold. This is particularly true when the PreTP is actually quite high or low.

There are some "tricks of the trade" to appreciate about using likelihood ratios. Tests (especially those with discriminating LRs) are best utilized when the PreTP is about 50%, as these will usually generate the widest movement on the Fagan monogram for PostTP determination (reader may use Figure 4 to confirm this). Overall, the higher the PreTP, the higher the PostTP, regardless of the LR value. In sequential testing, the PostTP of the previous test can be used as the PreTP for the next test(s), until a decision threshold has been crossed. Finally, it has been noted that most PreTP's tend to lie in the 10-90% range in clinical medicine, which likely define the most common decision thresholds (Grimes et al, 2005). The "rule of 15's" for positive tests suggested that for an LR+ = 2, the PostTP increases by 15%, for LR+ = 5, PostTP increases by 30%, and for LR+ = 10, PostTP increases by 30%. Inversely, the "reciprocal 2/5/10" rule for negative tests is as follows: the reciprocal of 2 = 0.5 (LR- = 0.5), then PostTP decreases by 15%; for reciprocal of 5 = 0.2 (LR- = 0.2), then PostTP decreases 30%, and finally if reciprocal of 10 = 0.1 (LR- = 0.1), then PostTP decrease 45% [7].

Recognizing Biases in Diagnostic Test Studies

As with any other published study design, those reporting results of diagnostic tests may be susceptible to various biases. The variable threats of different types of bias in diagnostic tests have been assessed previously [8,9]."Spectrum" bias occurs when the diagnostic test parameters discussed above are generated in one population, but applied in other populations with lesser utility. For example, diagnostic tests perform frequently better in patient populations with more serious or obvious manifestations of disease compared to those with milder manifestations or vague presentations [1,6]. "Selection" bias occurs in many study designs, when patients are not recruited consecutively, rather they are collected selectively, which may result in a non-representative population. "Verification" bias occurs when the results if the decision to perform the gold standard test is influenced by the results of the test being evaluated. In a diagnostic

test study, ideally all patients should receive the evaluation test and the gold standard test to allow for proper comparison. "Differential verification" refers to when one gold standard test is used to confirm a positive evaluative test, whereas a different gold standard (potentially inferior) may be used to confirm a negative evaluative test. For example, in the patient with low back pain, a physical exam suggestive of disc herniation may be confirmed with MRI, whereas an examination consistent with nonspecific mechanical low back pain may get a lumbar X-ray instead. The risk in this situation is that the verification strategy fails to identify all false-negative results. In keeping with differential verification bias is "partial verification bias," where not all patients are subjected to the reference standard test. The risk here is a preferential verification of positive evaluation results, leading to an overestimation of sensitivity and underestimation of specificity. "Incorporation" bias occurs when the gold standard reference test is interpreted with knowledge of the results of the evaluation test [10]. This lack of blinding inevitably results in overestimation of the evaluation test's diagnostic accuracy, especially if there is a subjective element to test interpretation. It would seem, however, that the average effects of inappropriate blinding are small [8]. "Publication bias" has been well described in therapeutic trials, outlining the tendency to publish positive trials compared to negative; this has been described for diagnostic test studies Table 4 [8].

Critical appraisal tools for diagnostic tests

It is up to the reviewers/editors of peer-reviewed journals and clinical readers to recognize the various biases and validity threats in any research manuscript, and evaluate how significant these threats are to their own clinical practice. It has been shown that critical appraisal skills are best acquired during undergraduate medical education, and less so at the resident level [11]. Although this review did not evaluate any critical appraisal education of post-graduate practicing clinicians, it is not unreasonable to assume that even less time is spent by busy clinicians learning critical appraisal skills. There are critical appraisal tools, however, that have been developed to help clinicians evaluate various study designs. For diagnostic tests, two installments of the User's Guide to the Medical Literature published in the Journal of the American Medical Association can be useful to guide readers as to the key elements of diagnostic test studies that should be optimized in order to maximize validity of results [12,13]. It is important to understand, however, that these quality "checklists" are qualitative in nature, asking a series of Yes/No/Unsure questions to various quality questions. The problem with such checklists is that it is not clear how many "Yes" or "No" or "Unsure" answers constitute small vs. larger validity threats, or if various quality components constitute more or less serious validity threats [14]. The most common tool currently used to evaluate quality of diagnostic test publications is the Quality Assessment of Diagnostic Accuracy Studies, (QUADAS tool) [15]. This tool is another qualitative checklist with no quantitative cutoffs [16] but has been validated by consensus. Uniform reporting standards also exist for publishing original diagnostic test studies (STARD criteria – Standards for Reporting Diagnostic Accuracy), although these are less likely important for clinical readers [17]. The advent of such reporting and critical appraisal tools has not necessarily translated to reliable evaluation of research manuscripts, however. There is evidence that the decision to publish manuscripts in peer-reviewed journals or to fund research proposals from large granting agencies may not correlate at all to quality evaluation checklist scores, but rather some other unreported features of the submissions [14]. Ultimately, it is the practicing clinician at the patient bedside who will make the final decision about using diagnostic test information in caring for their patients, and they should have the most unbiased information at their disposal to help aid this decision-making process.

Ordering diagnostic tests for patient reassurance?

Clinicians may feel pressured or obliged to order diagnostic tests that they know are not warranted by persistent patients who seek reassurance or at least validation for their symptomatology. This puts undue pressure on clinicians to inappropriately use resources in a non-evidence based manner, just to avoid complaints, litigation or any other potentially negative physician-patient interactions. For example, a recent survey of general practitioners found significant non-compliance with evidence-based clinical practice guidelines for managing low back pain, with 25% overuse of unnecessary imaging and inappropriate use of medications and activity recommendations [18]. Some reasons for guideline noncompliance included unyielding patient expectations regarding imaging, dissatisfaction with simple analgesics and variable understanding of patient education content. While it may seem expedient at the time to acquiesce to unreasonable patient diagnostic test requests in order to reduce patient anxiety, reduce original symptoms concerns or decrease future health care utilization, this does not actually happen. A recent systematic review in adult patients receiving reassurance testing with low probability of serious disease (14 trials, n=3828 patients, 2 low back pain studies included) has shown that there was NO overall reduction on patient illness worry, nonspecific anxiety, or long-term symptom persistence, although there was a small reduction in subsequent office visits [19]. The authors concluded that reassurance diagnostic testing is often low yield and resource-wasteful, and recommend adherence to best evidence-based practices when dealing with low risk patients. While this conversation may be difficult to have with some patients, it is nonetheless necessary from an evidence based management viewpoint.

Over-reliance on diagnostic tests?

There are times when a clinical history is very compelling, and you are very close to crossing a decision threshold to treat (T_{POS}), and you order a simple test to confirm your diagnosis. The problem with this approach is what to do if your test comes back negative and moves you back from T_{POS}, not past it? Sometimes the story is good enough to treat anyway. For example, a study of women with suspected urinary tract infections with negative dipstick tests found that those treated with antibiotics anyway had a mean reduction in symptoms by 4 days, with a number needed to treat = 4 [20]. In this case, the clinical assessment was sufficient to cross a threshold to initiate treatment, and avoid unnecessary urine testing delays. The same may apply to other pain conditions that are not likely to require diagnostic testing prior to initiating treatment plans. We recommend that each situation be considered carefully.

Application of Key Concepts to Key Reference Article

The results from the Chou review suggest a few predictors for persistent disabling low back pain, including maladaptive pain/fear avoidance behaviors, significant functional impairments, poorer

Table 4: Summary of biases in diagnostic test studies [9].

Type of Bias	Descriptor	Quantitative risk of bias*
Patient group factors: Accuracy of tests may vary between patient groups based on disease severity, comorbidities or alternative diagnoses		
- Cohort studies	- Cohort design; index test performed before reference standard	- 1.0
- Case control studies: severe cases vs. healthy controls	- Selection of severe cases and matched health controls	- 4.9 (0.6-37.3)
-Other case-control designs	- Case controls, avoidance of selecting patient at extreme ends of spectrum	- 1.1 (0.4-3.4)
- Selection: signs or symptoms	- Patient selection based on signs/symptoms of target condition	- 1.0
- Selection: referral for index test	- Patients selected based on those selected for index test	- 0.5 (0.3-0.9)
- Selection: other tests	- Patient selection based on other test results or referral for reference standard	- 0.9 (0.6-1.3)
- No limited challenge	- No additional criteria to exclude specific patients; risk of false negative or positive results	- 1.0
- Limited challenge	- Some additional criteria to exclude specific patients; risk of false negative or positive results	- 0.9 (0.6-1.3)
- Increased challenge	- Preferential inclusion of specific patients; risk of false negative or positive results	- 1.0 (0.6-1.7)
- Consecutive sampling	- Consecutive inclusion of patients satisfying inclusion criteria	- 1.0
- Nonconsecutive sampling	- Nonconsecutive inclusion of patients or selected cases	- 1.5 (1.0-2.1)
- Random sampling	- Inclusion of random subsample of patients meeting selection criteria	- 1.7 (0.9-3.2)
Verification procedures: ideally all index tests should be immediately verified with the reference standard, without intervening treatment		
- Same reference standard	- All results of index test confirmed with same reference standard	- 1.0
- Different reference standard	- Subset of index tests confirmed using a different reference standard	- 1.6 (0.9-2.9)
- Complete verification	- ALL index test results confirmed with reference standard	- 1.0
- Partial verification	- NOT ALL index test results confirmed with reference standard	- 1.1 (0.7-1.7)
- Single reference standard	- Reference standard is a single test/procedure	- 1.0
- Composite reference standard	- Reference standard is a combination of tests/procedures	- 0.9 (0.5-1.8)
- No incorporation	- Index test results NOT incorporated as part of reference standard	- 1.0
- Incorporation	- Index test results ARE incorporated as part of reference standard	- 1.4 (0.7-2.8)
- Time interval adequate	- Acceptable time window between index test and reference standard	- 1.0
- Time interval inadequate	- Unacceptable time window between index test and reference standard	- 1.1 (0.7-1.6)
- Treatment withheld	- No treatment given between index test and reference standard	- 1.0
- Treatment given	- Treatment given between index test and reference standard	- 0.9 (0.6-1.4)
Interpretation/reading: Knowledge of index test result prior to reading reference standard result, or vice versa, may enhance agreement		
- Double-blinded reading	-Interpretation of index test or reference standard without knowledge of other result	- 1.0
- Single/nonblinded reading	- Results of either test interpreted with prior knowledge of the other	- 1.1 (0.8-1.6)
Data Collection: Prospective data collection enables collection of higher quality data; retrospective data vulnerable to missing data or incomplete patient flow		
- Prospective data collection	- Data collection planned BEFORE performance of index test/reference standards	- 1.0
- Retrospective data collection	- Data collection planned AFTER index tests/reference standards done	- 1.6 (1.1-2.2)
Analysis: Data analysis choices may influence accuracy estimates, including cutoff choices for positive/negative, and exclusion of noninterpretable test results		
- Predefined or standard cutoff	- Positivity cutoff value for index test defined BEFORE data collection starts	-1.0
- Post hoc definition of cutoff	- Positivity cutoff value for index test defined AFTER data collection	- 1.3 (0.8-1.9)
- Noninterpretable results reported	- Explicit reporting of indeterminate/ noninterpretable test results & outliers	- unable to calculate due to incomplete reporting
- Noninterpretable results not reported	- Indeterminate/noninterpretable results & outliers not reported	
- No dropouts	- Data on >90% of included patients available for analysis	- unable to calculate due to incomplete reporting
- Dropouts	- Data on <90% included patients available for analysis	

*Quantitative risk of bias represented as Relative Diagnostic Odds Ratio (RDOR; 95%CI). Any value of RDOR >1.0 suggests that there is exaggeration of diagnostic accuracy based on design deficiencies based on that bias item.

general health status, and the presence of psychiatric comorobities (Table 1). However, the likelihood ratios associated with these predictor variables are generally weak (especially LR+), ranging from 0.33-2.5, which is hardly compelling to rule in likely disability, regardless of high the PreTP is (reader can confirm this on Fagan nomogram). The authors rightly state limitations of individual studies including variable risk factor and low back outcome definitions, small numbers of studies/enrolled patients, no subgroup analyses of back pain etiologies. Based on these limitations, the authors advocate for less dependence on individual risk factors and more for use of standardized risk prediction instruments that have some proven reliability (e.g. Roland Morris Questionnaire, Oswestry Disability Index, etc.). Finally, the authors suggest adherence to best evidence-based clinical practice guidelines for managing such patients.

Scenario resolution

You make a preliminary diagnosis of non-specific mechanical LBP, and educate/reassure the patient of this diagnosis. In adherence with the Alberta TOP guideline, you advise ongoing activity, return to work if possible, alternating cold/heat packs and a medication course of acetaminophen or over-the-counter non-steroidal anti-inflammatory. The patient has coverage for acupuncture and chiropractic manipulation, so you encourage them to pursue these treatment modalities as needed. You plan to reassess the patient again in 4-6 weeks time, to evaluate progress and recovery. You also ask the patient to start using the Roland Morris Questionnaire to track weekly progress, which the patient agrees to do. At the follow up visit, you can reassess progress in activity, pain resolution and possible need for imaging.

Conclusion

Busy clinicians face a myriad of decisions every day in their patient care encounters. Part of the assessment of new patient conditions is the judicious use of diagnostic testing when a decision to manage or not manage a disease of interest is not readily reached after clinical assessment alone. The use of diagnostic tests to cross decision thresholds is appropriate, as long as they are used for the purposes of changing management and not just fishing for non-useful information. Understanding the diagnostic test characteristics, appropriate populations to test, and limitations of diagnostic tests can lead to more appropriate testing choices. When reading about new diagnostic test studies, educated readers should watch for different types of biases in such studies, and have appropriate critical appraisal skills to detect them. Ordering tests for non-management a reason (e.g. patient stress and anxiety reduction, over-reliance to confirm that which you already know) are not warranted, needlessly generates resource costs and should be avoided. Well constructed evidence-based clinical practice guidelines appropriate for your patient populations are likely your best tools for managing various clinical conditions.

References

1. Worster A, Innes G, Abu-Laban RB. Diagnostic testing: an emergency medicine perspective. See comment in PubMed Commons below CJEM. 2002; 4: 348-354.

2. Alberta Toward Optimized Practice, Low Back Pain. www.topalbertadoctors.org, 2009.

3. Chou R, Shekelle P. Will this patient develop persistent disabling low back

pain? See comment in Pub Med Commons below JAMA. 2010; 303: 1295-1302.

4. Emery DJ, Shojania KG, Forster AJ, Mojaverian N, Feasby TE. Overuse of magnetic resonance imaging. See comment in PubMed Commons below JAMA Intern Med. 2013; 173: 823-825.

5. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. See comment in PubMed Commons below CJEM. 2006; 8: 19-20.

6. Montori VM, Wyer P, Newman TB, Keitz S, Guyatt G. Evidence-Based Medicine Teaching Tips Working Group . Tips for learners of evidence-based medicine: 5. The effect of spectrum of disease on the performance of diagnostic tests. See comment in PubMed Commons below CMAJ. 2005; 173: 385-390.

7. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. See comment in Pub Med Commons below Lancet. 2005; 365: 1500-1505.

8. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. See comment in PubMed Commons below JAMA. 1999; 282: 1061-1066.

9. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. See comment in PubMed Commons below CMAJ. 2006; 174: 469-476.

10. Worster A, Carpenter C. Incorporation bias in studies of diagnostic tests: how to avoid being biased about bias. See comment in Pub Med Commons below CJEM. 2008; 10: 174-175.

11. Norman GR, Shannon SI. Effectiveness of instruction in critical appraisal (evidence-based medicine) skills: a critical appraisal. See comment in PubMed Commons below CMAJ. 1998; 158: 177-181.

12. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. See comment in PubMed Commons below JAMA. 1994; 271: 389-391.

13. 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? The Evidence-Based Medicine Working Group. See comment in PubMed Commons below JAMA. 1994; 271: 703-707.

14. Upadhye S. Brian Rowe, Eddy S Lang, Michael D Brown, Debra Houry, David H Newman, Peter C Wyer, editors. Pitfalls in critical appraisal. Chapter 3. In: Evidence-Based Emergency Medicine Book ISBN: 978-1-4051-6143-5 BMJ Books November 2008.

15. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. See comment in Pub Med Commons below BMC Med Res Methodol. 2006; 6: 9.

16. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. See comment in Pub Med Commons below BMC Med Res Methodol. 2003; 3: 25.

17. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem. 2003; 49: 7-18.

18. Williams CM, Maher CG, Hancock MJ, McAuley JH, McLachlan AJ, Britt H, et al. Low back pain and best practice care: A survey of general practice physicians. See comment in PubMed Commons below Arch Intern Med. 2010; 170: 271-277.

19. Rolfe A, Burton C. Reassurance after diagnostic testing with a low pretest probability of serious disease. JAMA Intern Med. 2013; 173: 407-416.

20. Richards D, Toop L, Chambers S, Fletcher L. Response to antibiotics of women with symptoms of urinary tract infection but negative dipstick urine test results: double blind randomized controlled trial. BMJ. 2005; 331:143.

Phys Med Rehabil Int - Volume 1 Issue 1 - 2014
ISSN : 2471-0377 | www.austinpublishinggroup.com

Citation: Upadhye S and Kumbhare DA. Diagnostic Test Principles: Overview and Applications for Pain Medicine. Phys Med Rehabil Int. 2014;1(1): 8.