Diagnostic testing: an emergency medicine perspective

Andrew Worster, MD, MSc(HRM);* Grant Innes, MD;† Riyad B. Abu-Laban, MD, MHSc‡

ABSTRACT

Emergency physicians use diagnostic tests extensively, and the ability to order and interpret test results appropriately is a critical skill. An understanding of sensitivity, specificity, predictive values and likelihood ratios, as well as an awareness of the importance of pre-test probability, is essential. The purpose of this article is to explain, in a straightforward and clinically applicable manner, the core concepts related to diagnostic testing.

Key words: diagnosis, sensitivity, specificity, predictive value, likelihood ratio, prevalence, emergency medicine

RÉSUMÉ

Les médecins d'urgence utilisent beaucoup les épreuves diagnostiques et il est essentiel qu'ils sachent interpréter correctement les résultats. Une compréhension de la sensibilité, de la spécificité, des valeurs prédictives et des rapports de probabilité, ainsi qu'une conscience de l'importance des probabilités pré-tests, est primordiale. Le présent article a pour but d'expliquer, de façon claire et applicable en pratique clinique, les concepts de base liés aux épreuves diagnostiques.

Introduction

Diagnostic tests are used to categorize patients, to ascertain disease severity, to prognosticate, to assess response to treatments and, most importantly, to help establish diagnoses. Unfortunately, most tests don't "make" diagnoses; they supplement clinical judgement and reduce the level of diagnostic uncertainty. Unless applied and interpreted carefully, tests can be misleading.

The basic premise of diagnostic testing is that there are 2 populations of people — those with the disease in question and those without — who differ on at least one testable parameter. For example, patients with pneumonia have infiltrates on x-ray, while those without pneumonia do not. The real world, however, is not so simple. Not everyone with pneumonia has an infiltrate and not everyone with an infiltrate has pneumonia. Patient variability and test variability result in an overlap between the results for diseased and normal populations for virtually all tests (Fig. 1).

Test variability may be related to the test, the interpreter, or both. Test results may vary depending on the duration of symptoms (e.g., troponin level in myocardial infarction) or the stage of the illness (e.g., lipase levels in acute and chronic pancreatitis). Test results may differ because of lab equipment, reagents, procedure, or even lab error. Many tests, like ECGs and imaging studies, require interpretation that is subject to variability. The interpretation of tests may be biased by prior test results (e.g., awareness of abnormal cardiac marker levels may modify the interpretation of an ECG) or by clinical information (e.g., knowledge that a patient has fever, dyspnea and rust-coloured sputum may influence the interpretation of a borderline chest x-ray). For all these reasons, test results cannot always be accepted at

From the *Division of Emergency Medicine, Hamilton Health Sciences and McMaster University, Hamilton, Ont.; the †Department of Emergency Medicine, Providence Health Care and St. Paul's Hospital, Vancouver, BC; the ‡Department of Emergency Medicine, Vancouver General Hospital, Vancouver, BC; and the †‡University of British Columbia, Vancouver, BC

Received: Apr. 15, 2002; final submission: June 15, 2002; accepted: June 20, 2002

This article has been peer reviewed.

face value, and tests cannot be interpreted without considering pretest (clinical) probability of disease.

Measures of test efficacy

Most objective tests assess a measurable parameter and classify the patient as "normal" or "abnormal." "Normal" is typically established by determining test values in disease-free people and identifying the range in which 95% (2 standard deviations) of this population lies. Using this definition, it is apparent that 5% of disease-free people will have values outside the normal range; in fact, for virtually all tests, abnormal values occur in patients without disease and normal values occur in patients with disease. In addition, because physiologic parameters vary from patient to patient (just like weight and height), it is sometimes more useful to know normal values for a patient than to know normal values for the general population. Occasionally a changing value, which is abnormal for the individual but still within the normal range, may signify disease.

Test results can be categorized as true-positive, false-positive, false-negative or true-negative, relative to a reference (gold) standard that correctly identifies patients with and without disease. Unfortunately, we have few good gold standards and often use surrogates (e.g., laparotomy is the real gold standard for gallstones, and ultrasound is the surrogate we generally use). For many conditions there is no true gold standard test, and clinical outcome may be the best standard available. Table 1, a standard 2×2 table, relates true disease status to diagnostic test result and allows us to calculate all of the important test parameters. Understanding the 2×2 table is key to understanding the various terms discussed below, which quantify the performance of diagnostic tests.

Sensitivity, specificity and accuracy

Sensitivity, or true-positive rate, refers to the probability that a test will be positive in patients known to have the disease.[1] Highly sensitive tests have few false-negative results and are most useful to rule out disease. If a test is 96% sensitive, this means 96% of patients with the target disorder will have a true-positive result and 4% will have a false-negative result.[2] We use highly sensitive tests when we need to rule out dangerous conditions (e.g., lumbar puncture for subarachnoid hemorrhage).

Specificity, or true-negative rate, refers to the probability that a test will be negative in patients known to be disease free. Highly specific tests have few false-positive results and are most useful to rule in disease.[1] If a test is 96% specific, this means 96% of patients without the target disorder will have a true-negative result and 4% will have a false-positive result.[2] Specific tests are important in situations where a false-positive test could lead to harm, for example when the therapy is potentially dangerous (e.g., long-term anticoagulation).

Accuracy describes the overall test performance in patients with and without disease, and is calculated by determining the total number of true-positive and true-negative results, then dividing by the total number of tests done. Although this term is often used, it has little clinical relevance because the more important question is how the test performs for the clinical "mode" a physician is in. When physicians order a test, they are usually in either a "rule-in" or a "rule-out" mode (and only occasionally in both). A mnemonic that is helpful in selecting tests appropriate for

Table 1. Diagnostic test parameters summarized by a 2 × 2 table

		Actual patient status (truth)		
		Disease present	**Disease absent**	**Total no. of patients**
Test result	**Positive**	True positive (A)	False positive (B)	With positive test (A+B) → Positive predictive value = A / A+B
	Negative	False negative (C)	True negative (D)	With negative test (C+D) → Negative predictive value = D / C+D
	Total no. of patients	With disorder (A+C) → Sensitivity = A / A+C	Without disorder (B+D) → Specificity = D / B+D	(A+B+C+D) → Accuracy = A+D / A+B+C+D

the clinical mode is the rule of "spin and snout." Highly specific tests rule in disease (Sp-In), while highly sensitive tests rule out disease (Sn-Out).[1]

Unfortunately, diagnostic tests rarely provide binary (positive vs. negative) results. Most, such as white blood count, creatinine, and troponin, are reported as continuous values. Even pregnancy testing, which we think of as positive or negative, is actually based on a β-hCG (beta-human chorionic gonadotropin) threshold chosen to best distinguish pregnant from nonpregnant populations. When the test result is a numeric value, changing the "cut-off" threshold will change sensitivity and specificity, thus changing the ability of the test to rule in or rule out disease. If the cut-off is lowered, the test will become more sensitive and less specific (fewer false negatives, more false positives, so better at ruling out disease). If the cut-off is raised, the test will become less sensitive and more specific (more false negatives, fewer false positives, so better at ruling in disease). This trade-off between sensitivity and specificity exists for virtually all tests, and manipulating the cut-off level can usually only improve one parameter at the expense of the other.[2,3] Receiver operating characteristic (ROC) curves can be used to select optimal cut-off values and determine the diagnostic performance of a test across a range of values. A detailed discussion of ROC curves is beyond the scope of this article, however these are described elsewhere in the emergency medicine literature.[4]

Positive and negative predictive values

Sensitivity and specificity describe how tests perform in people who are known to have or not have the disease in question. But if we knew the patient's true disease status before testing, we wouldn't need to do a test![5] In clinical medicine, it is more common to have a patient with an unknown disease status, and then to be faced with interpreting a test result. Given the test result, we need to know how likely it is that the patient has or does not have the disease in question. Predictive values provide this information. Positive predictive value (PPV) tells us the probability of disease if the patient's test is positive, while negative predictive value (NPV) tells us the probability that the patient is disease-free if the test is negative.[6]

Unlike sensitivity and specificity, which are generally considered stable characteristics of a diagnostic test, the predictive value of a test may vary dramatically depending on the pretest probability of disease in the patient being tested. Pretest probability, also known as prevalence, describes the clinical likelihood — before doing a test — that the patient has the target disease.

When interpreting test results, it is important to remember that a test's PPV is better in high prevalence populations, while NPV is better in low prevalence populations. For example, highly sensitive rapid HIV tests have very good PPV when used on a population of patients with HIV risk factors and opportunistic infections, but the same tests have much lower PPV when used indiscriminately for the approval of life insurance coverage. In this latter situation, the majority of positive tests will be false-positive. It is clear that, while predictive values are more clinically relevant than sensitivity and specificity because they help us interpret test results in patients with differing pretest probability of disease, tests must be used selectively on appropriate patient populations and interpreted differently depending on disease prevalence (pretest likelihood).

Predictive values taken in isolation can be misleading; they do not tell us everything we need to know about the diagnostic utility of a test. To illustrate, studies of diagnostic tests often report excellent NPV, without similarly highlighting PPV. Many physicians believe that if a test has good NPV it can be trusted to rule out disease, but as the following scenario illustrates, this assumption can be wrong.

Scenario: A colleague tells you that that he no longer uses cardiac marker assays because the NPV of a coin toss is just as good. Intrigued, you apply a coin toss (heads = positive; tails = negative) to your next 100 emergency department (ED) patients presenting with chest pain. Table 2 shows that, in the group of 50 patients whose toss came up tails, there were only 3 patients with myocardial infarction (MI) (NPV = 94%), and in the group of 50 patients whose toss came up heads, there were 3 patients with MI (PPV = 6%). You conclude that the coin toss is a poor positive predictor but an excellent negative predictor, and you present this information at cardiac care unit (CCU) rounds the following week. To your dismay, the cardiologists have conducted a similar experiment in the CCU and derived conflicting results. Table 3 shows that, in the group of 50 CCU patients whose coin toss came up tails, there were 45 patients with MI (NPV = 5%), and in the group of 50 patients whose coin toss came up heads, there were 45 patients with MI (PPV = 90%). The cardiologists concluded that the coin toss has poor NPV and excellent PPV, and intend to publish their results.

Table 2. Predictive value of the coin toss scenario in ED patients with chest pain (MI prevalence = 6%)

	AMI	No AMI	
Heads (+)	3	47	50
Tails (–)	3	47	50
	6	94	100

Sensitivity = A / A+C = 3 / (3 + 3) = 50%
Specificity = D / B+D = 47 / (47 + 47) = 50%
Positive predictive value = A / A+B = 3 / 3 + 47) = 6%
Negative predictive value = D / C+D = 47 / (3 + 47) = 94%
AMI = acute myocardial infarction

Clearly both conclusions are wrong. This scenario demonstrates that the coin toss is a useless test, that sensitivity (50%) and specificity (50%) remain constant regardless of pretest probability, and that the predictive value of a test may differ dramatically depending on prevalence of disease in the population being tested. In a low prevalence population like the ED, where 6% of patients had MI, positive tests were usually wrong (poor PPV), but in a high prevalence population like the CCU, where 90% of patients had MI, negative tests were usually wrong (poor NPV). This is the reason we were taught to "treat the patient, not the test," and this is why test results must always be interpreted in light of pretest probability. Predictive values reported in published studies can only be generalized to our clinical practice if the disease prevalence (pretest probability) is similar in the study setting and our clinical practice. This scenario also illustrates why it is misleading to publish NPV without PPV, or to report predictive values without also reporting sensitivity and specificity.

Table 3. Predictive value of the coin toss scenario in CCU patients with chest pain (MI prevalence = 90%)

	AMI	No AMI	
Heads (+)	45	5	50
Tails (–)	45	5	50
	90	10	100

Sensitivity = A / A+C = 45 / (45 + 45) = 50%
Specificity = D / B+D = 5 / (5 + 5) = 50%
Positive predictive value = A / A+B = 45 / 45 + 5) = 90%
Negative predictive value = D / C+D = 5 / (45 + 5) = 10%
CCU = cardiac care unit

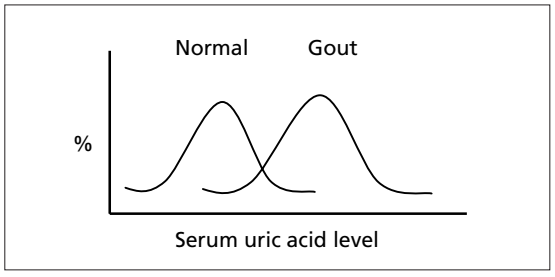

Fig. 1. Uric acid levels in healthy and diseased patients. There is variability in the normal and in the diseased population, and overlap between the two groups. Some uric acid levels are therefore compatible with health or disease.

Decision thresholds and diagnostic tests

Perhaps the most effective way to use diagnostic tests is to visualize every patient on a clinical decision line (Fig. 2). The left end of the line represents a pretest probability of zero — absolute clinical certainty that the patient does not have the disease in question — and the right end of the line represents a probability of 100% — absolute clinical certainty that the patient has the disease in question. Most patients, based on a clinical assessment of risk factors, history and physical findings, will fall somewhere between these 2 points.

The "negative decision threshold" (T_0) is the point below which pretest disease probability, or level of diagnostic suspicion, is low enough to allow a negative treatment decision (e.g., discharge) without further investigation. The "positive decision threshold" (T_1) is the point above which the pretest probability, or level of diagnostic suspicion, is high enough to justify a positive treatment decision (e.g., surgery) without further testing. Sometimes after clinical assessment alone, diagnostic suspicion will be beyond a T_0 or T_1 point (e.g., obvious chest wall pain or strongly suspected appendicitis), in which case further testing is unnecessary because it will not influence treatment.

Decision thresholds are not static; they vary with the disease in question. In situations where it would be disastrous to miss the diagnosis (e.g., meningitis), a sensitive test is necessary, so decision thresholds should be set low and test cut-offs adjusted for maximum sensitivity. If missing the diagnosis is unlikely to cause harm (e.g., cholelithiasis), but treating the suspected condition could cause harm (e.g., laparotomy), a specific test is necessary, so decision

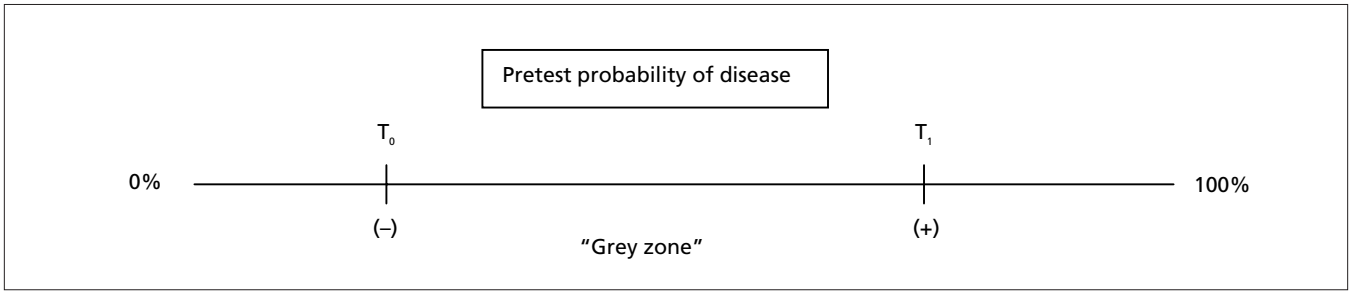

Fig. 2. A Clinical Decision Line. (–) T_0 represents the negative decision threshold, a level of diagnostic suspicion below which the diagnosis is "ruled out" on clinical grounds without further investigations. (+) T_1 represents the positive decision threshold, a level of diagnostic suspicion above which treatment is warranted without the need for more expensive or invasive tests.

thresholds should be set higher and test cut-offs adjusted for maximum specificity.

The zone between the 2 decision thresholds is the clinical "gray-zone," an area of uncertainty where further diagnostic testing is required. Often clinicians will find themselves in this zone after an initial history and physical exam (e.g., atypical chest pain compatible with pulmonary embolism), and sometimes they will still be in this zone after a non-definitive test results is obtained (e.g., an indeterminate probability VQ [ventilation perfusion] scan). The main role of a diagnostic test is to carry us across a negative or positive decision threshold. For patients who are close to a decision threshold, a "weak" test may suffice, but for those who are far from a decision threshold, a "strong" test is required.

Likelihood ratios

Likelihood ratios (LRs) are the most useful single indicator of a test's diagnostic strength, therefore of the degree to which it can modify pretest probability and facilitate clinical decision-making. Positive LR (LR+) is the ratio of true-positive rate to false-positive rate, while the negative LR (LR–) is the ratio of the false-negative rate to true-negative rate. LRs are calculated using these formulae:

$$\text{Positive LR (LR+)} = \text{Sensitivity} \;/\; (1 - \text{Specificity})$$
$$\text{Negative LR (LR–)} = (1 - \text{Sensitivity}) \;/\; \text{Specificity}$$

As LR+ increases, the test becomes a stronger positive predictor, and as LR– decreases, the test becomes a stronger negative predictor. Positive LRs between 1.0 and 3.0 are very weak, and those greater than 10 generate large and often conclusive changes in post-test probability. LRs greater than 20 are usually diagnostic. Conversely, negative LRs between 0.3 and 1.0 are relatively weak, and those less than 0.1 generate large and often conclusive changes in post-test probability. Negative LRs less than 0.05 are usually diagnostic. To rule out disease, a sensitive test (low LR–) is required, whereas to rule in disease, a specific test (high LR+) is required. In the coin-toss scenario above, the LR+ and the LR– are both 1.0, indicating that the coin toss is a useless test.

LRs are less intuitive than other test parameters, and physicians may erroneously believe that LRs represent the factor by which pretest probability is multiplied to get post-test probability. In reality, LRs are the multiplicative factor linking pre- and post-test "odds." But odds are a concept that only seasoned gamblers understand (odds = probability / 1 – probability). To illustrate, if a team has an 80% chance of winning a game, their odds of winning are 4 to 1. Because LRs are based on odds, a nomogram must

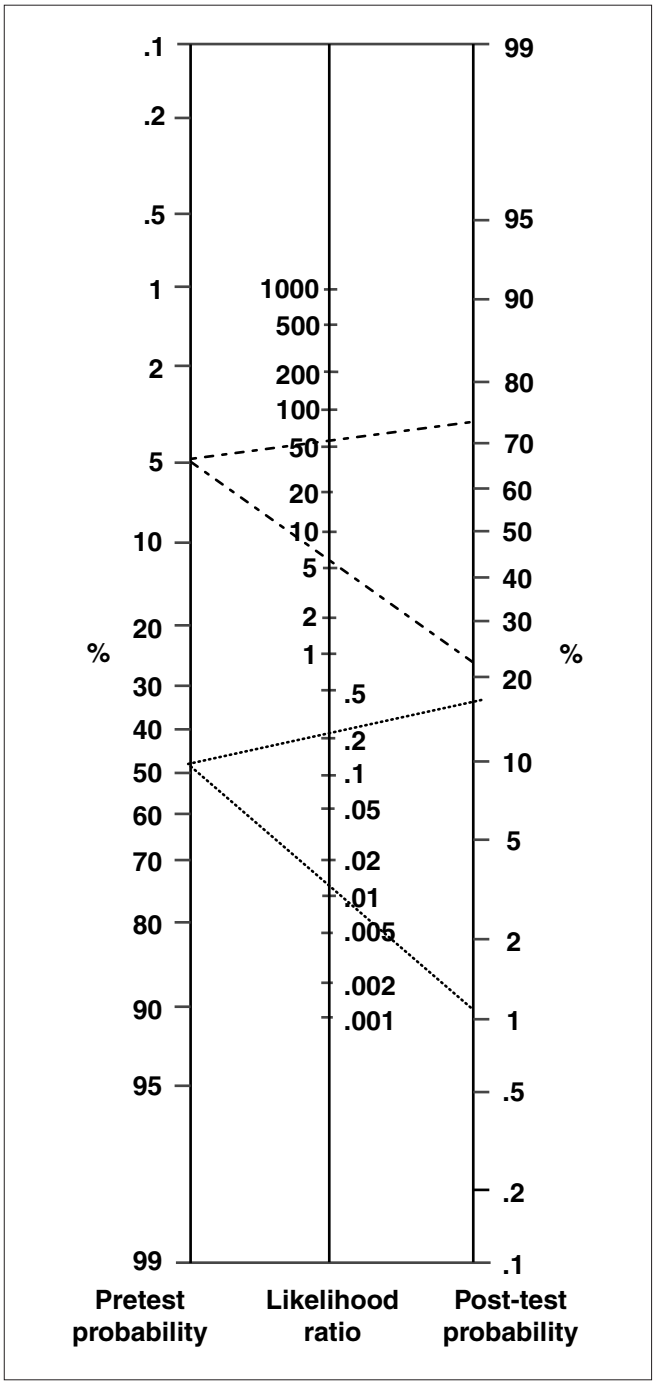

Fig. 3. This illustrates the use of the Fagan nomogram and likelihood ratios to determine post-test probability of disease. Patient A (– – – – – – –) has a 5% pretest probability of disease. If a "weak" test (LR+ = 5) comes back positive, post-test probability increases to 21%. If a strong test (LR+ = 50) comes back positive, it increases it to 72%. Patient B (.....) has a 50% pretest probability of disease. If a weak test (LR– = 0.2) comes back negative, post-test probability decreases to 17%. If a strong test (LR– = 0.01) comes back negative, it decreases it to 1%. Adapted with permission from Fagan TJ. Nomogram for Bayes's theorem [letter]. *N Engl J Med* 1975;293(5):257.

be utilized when converting pretest probability to post-test probability using LRs (Fig. 3).[8] Table 4 illustrates LRs for some common diagnostic tests, and these values, in combination with the nomogram, allow determination of post-test probability and whether a positive or negative decision threshold has been crossed.[9]

Scenario: After examining a patient with vague abdominal pain and distension, you believe there is a 20% pretest probability of bowel obstruction. The subsequent x-ray shows definite air–fluid levels. Knowing that abdominal x-rays have an LR+ of 10 in the setting of bowel obstruction (Table 4), you use the Fagan nomogram (Fig. 3) to estimate that the post-test probability of bowel obstruction is 70%. (This is done by drawing a line linking the pretest probability, 20%, through an LR+ of 10, to obtain post-test probability.) This probability change, from 20% to 70% is sufficient to carry you across a positive decision threshold and arrange hospitalization.

Spectrum bias

Most diseases have a "spectrum" of possible presentations or stages. Spectrum bias exists when test performance parameters are misrepresented for a given situation because they were measured on a different spectrum of patients than the test is now being applied to. It explains the fact that tests perform better in patients with severe or advanced disease than in those with subtle or early disease. For example, new generation CT scanners are reported to be over 95% sensitive for detecting subarachnoid hemorrhage. However, if we eliminate clinically obvious cases with massive hemorrhage, and consider only patients with sentinel bleeds and vague findings (the spectrum of patients who most need a diagnostic test), CT sensitivity falls substantially. Spectrum bias occurs with many diagnostic tests. For example the white blood cell count is more likely to be elevated in patients with ruptured appendicitis, peritonitis and fever than in those with minimal early findings. Similarly, cardiac markers are more likely to be elevated in hypotensive patients with obvious ST-elevation than in subtle patients who have nondiagnostic ECGs. Spectrum bias explains why diagnostic tests tend to "miss" the same patients that clinicians do, and why tests tend to perform the worst in patients whose diagnosis is not clinically obvious.

Summary and key points

Sensitivity and specificity are widely understood, but have limited clinical relevance and are often inappropriately emphasized. Predictive values are more clinically relevant and useful than sensitivity and specificity, because they relate to "real world" problem of determining the probability of disease in an individual patient, and because they reflect the important influence of pretest probability (prevalence) on diagnostic test performance. Predictive values can be misleading: poor tests can appear to have excellent PPV or NPV if they are used in high- or low-prevalence popula-

Table 4. Positive (LR+) and negative (LR–) likelihood ratios for commonly used tests[11,12]

Test	Suspected diagnosis	LR+	LR–
X-rays			
Abdominal	Bowel obstruction	10.0	0.0
Bone	Osteomyelitis	5.6	0.55
Chest	Pulmonary embolism	1.7	0.84
Endoscopy	Peptic ulcer	100	0.05
Lumbar puncture (>5 WBC)	Meningitis	50	0.01
Lumbar puncture	SAH	5.0	0.0
Head CT scan	SAH (1 to 5 days)	17	0.16
Head CT scan	SAH (>5 days)	10	0.53
ECG (single)	Acute myocardial infarction	30	0.44
ECG (serial)	Acute myocardial infarction	68	0.32
Compression ultrasound	Proximal deep vein thrombosis	19	0.05
Ultrasound	Gallbladder stones	18	0.15
Leukocyte esterase	Urinary tract infection	20	0.15
Urine micro exam	Urinary tract infection	90	0.10
White blood count	Appendicitis	2.2	0.18
Noncontrast helical CT	Acute urolithiasis	23	0.05
Urography	Acute urolithiasis	9	0.33

WBC = white blood count; SAH = subarachnoid hemorrhage; CT = computed tomography; ECG = electrocardiography

tions, respectively, and reported predictive values are often more reflective of pretest probability (disease prevalence) than they are of intrinsic test characteristics. This is the reason we were taught to "treat the patient, not the test," and the reason why test results must always be interpreted in light of pretest probability.

LRs, while less intuitive and more poorly understood, are the best overall indicators of the diagnostic strength of tests and deserve greater physician awareness. Visualizing a clinical decision line, and consciously considering whether one is in a rule-in or rule-out mode, can be helpful in test ordering and interpretation. Finally, the possibility of spectrum bias should always be considered before generalizing reported test performance to the ED population.

Competing interests: None declared.

References

1. Sackett DL, Straus S. On some clinically useful measures of the accuracy of diagnostic tests [editorial]. ACP J Club 1998;129: A17-9.

2. Nettleman MD. Receiver operator characteristic (ROC) curves. Infect Control Hosp Epidemiol 1988;9:374-7.

3. Shapiro DE. The interpretation of diagnostic tests. Stat Methods Med Research 1999;8:113-34.

4. Grzybowski M, Younger JG. Statistical methodology: III. Receiver operating characteristic (ROC) curves. Acad Emerg Med 1997;4(8):818-26.

5. Gallagher EJ. Clinical utility of likelihood ratios. Ann Emerg Med 1998;31:391-7.

6. Young MJ, Fried LS, Eisenberg JM, Hershey JC, Williams SV. The single-cutoff trap: implications for Bayesian analysis of stress electrocardiograms. Med Decision Making 1989;9:176-80.

7. 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. JAMA 1994;271(9):703-7.

8. Fagan TJ. Nomogram for Bayes's theorem [letter]. N Engl J Med 1975;293(5):257.

9. Sox HC Jr, Blatt MA, Higgins MC, Marton KI. Medical decision making. Boston: Butterworths; 1988.

Correspondence to: Dr. Andrew Worster, Department of Emergency Medicine, Hamilton Health Sciences, 237 Barton St. E, Hamilton ON L8N 3Z5; fax 905 527-7051, aworster@rogers.com