Diagnosis: Making the Best Use of Medical Data

MARK H. EBELL, MD, MS, *University of Georgia, Athens, Georgia*

To take the best possible care of patients, physicians must understand the basic principles of diagnostic test interpretation. Pretest probability is an important factor in interpreting test results. Some tests are useful for ruling in disease when positive or ruling out disease when negative, but not necessarily both. Many tests are of little value for diagnosing disease, and tests should be ordered only when the results are likely to lead to improved patient-oriented outcomes. (*Am Fam Physician.* 2009;79(6):478-480. Copyright © 2009 American Academy of Family Physicians.)

This is the third article in a six-part series about finding evidence and putting it into practice

Although evidence-based medicine is often associated with randomized controlled trials and treatment decisions, the past 20 years have seen an explosion in our knowledge about diagnosis. New tests, such as the brain natriuretic peptide (BNP) and D-dimer tests, have been developed, and physicians have better data on older tests and on the history and physical examination.

Adopting New Tests

New tests are usually described in terms of their sensitivity and specificity. A sensitive test is good for detecting disease when it is present, whereas a specific test is good for identifying the absence of disease in healthy patients. But there are several other important factors that make a test worth adopting, including cost, availability, and the potential for harm. Most importantly, does the information help physicians take better care of patients and improve patient-oriented outcomes? Knowing with greater certainty that a patient has a disease is helpful only if this knowledge leads to an improvement in treatment that increases how long or how well the patient lives. Tests can be harmful when they lead to unnecessary invasive procedures or unneeded worry. For example, if an older patient who smokes presents with dyspnea of uncertain origin, the physician might consider electrocardiography (ECG), echocardiography, radiography, and BNP measurement. Should all four tests be ordered? Which ones merely add cost without improving patient-oriented outcomes? In this case, a study in several European emergency departments found that use of the BNP test in the setting described above reduced the length of hospitalization and saved money.[1] Although chest radiography and ECG probably should be ordered, an echocardiogram isn't necessary if the BNP levels are normal.

Knowing the sensitivity and specificity of tests is useful to researchers, but it is the source of much frustration to physicians because these numbers don't describe the test from our perspective. Sensitivity and specificity tell us the likelihood of a positive or negative test, given that the patient does or does not have the disease in question. Of course, if we knew whether or not the patient had the disease, we wouldn't need the test!

Knowing the predictive values and post-test probabilities is more helpful because these values answer the following key questions: (1) if a test is positive, what is the likelihood of disease (positive predictive value or post-test probability of a positive test)? and (2) if a test is negative, how likely is it the patient does not have the disease (negative predictive value or post-test probability of a negative test)?

What does this mean to you as a physician? First, always consider whether the information gained from the test is likely to improve patient-oriented outcomes. Second, think in terms of predictive value. How much does a positive test increase the likelihood of disease, and how much does a negative test decrease it?

Discontinuing Tests

Some tests that were once thought to be helpful turn out to be inaccurate when

Table 1. Tests and Findings with Little or No Diagnostic Value

Diagnosis	Test or finding	Sensitivity (%)	Specificity (%)	LR+	LR−
Acute cholecystitis[2]	Elevated alanine transaminase or aspartate transaminase level	38	62	1.0	1.0
Breast cancer (patient with spontaneous single-duct nipple discharge)[3]	Ultrasonography	36	68	1.1	0.94
Iron deficiency anemia[4]	Mean corpuscular volume of 75 to 79 μm³ (75 to 79 fL)*	—	—	1.0	—
Lumbar spinal stenosis[5]	Pain is worse with walking	71	30	1.0	1.0
Migraine headache[6]	Headache is triggered by menses	44	56	1.0	1.0
Ovarian cancer[7]	Indigestion	36	63	1.0	1.0
Peripheral artery disease[8]	Weak femoral artery pulse	33	67	1.0	1.0
Pulmonary embolism[9]	Ventilation-perfusion scanning (intermediate probability)*	—	—	1.2	—

LR+ = positive likelihood ratio; LR− = negative likelihood ratio.

*—Tests with no single cutoff or cut-point.

Information from references 2 through 9.

carefully studied *(Table 1)*.[2-9] Positive and negative likelihood ratios (LRs) tell us the extent to which a positive or negative test increases or decreases the likelihood of disease. LRs greater than 5.0 to 10.0 significantly increase the likelihood of disease, and those less than 0.1 to 0.2 significantly decrease it. LRs between 0.2 and 5.0 change the likelihood of disease much less, especially as they approach 1.0. Although the tests listed in *Table 1* are widely taught and widely used, their LRs are close to 1.0; therefore, they have little or no value for diagnosis.[2-9]

Some tests have no single cutoff or cut-point, such as yes or no. Instead, they can have a range of values and a range of LRs *(Table 2)*.[10] This type of LR gives us the most information from a test result.

Table 2. Likelihood of Endometrial Cancer in Women with Postmenopausal Vaginal Bleeding

Thickness of endometrial stripe (mm)	Likelihood ratio	Post-test probability (%)*
≤ 4	0.02	0.2
5	0.21	2.3
6 to 10	0.5	5.3
11 to 15	2.2	19.6
16 to 20	6.4	41.6
21 to 25	9.0	50.0
> 25	15.2	62.8

*—Assumes a pretest probability of 10 percent in women with postmenopausal vaginal bleeding.

Information from reference 10.

Ruling In and Ruling Out Disease

Some tests are good at ruling in disease when the results are positive, but they do not rule out disease when they are negative (or vice versa). This can be confusing to physicians who think that tests behave symmetrically (i.e., they are equally good at ruling in and ruling out disease). Tests that are useful only for ruling in disease tend to have a sensitivity near 50 percent, but a very high specificity. Conversely, tests that are useful only for ruling out disease have a very high sensitivity, but a modest specificity. A good example comes from a meta-analysis of D-dimer testing in patients with suspected pulmonary embolism.[11] A rapid D-dimer test result of greater than 500 mcg per L (2.74 nmol per L) was 99 percent sensitive, but only 44 percent specific for diagnosis of pulmonary embolism. This corresponds to positive and negative LRs of 1.8 and 0.2, respectively. An online clinical calculator (http://www.dokterrutten.nl/collega/LRcalcul.html) shows that if a patient has a 10 percent pretest probability of pulmonary embolism, that probability increases to 17 percent if the D-dimer results are abnormal (not clinically helpful). However, if the D-dimer results are normal, the probability decreases to only 0.2 percent. Thus, this test is very good at ruling out pulmonary embolism when negative in a low-risk patient, but it is of little value for ruling in pulmonary embolism when results are abnormal in the same patient.

Interpreting Test Results

A common misconception is that evidence-based medicine and practice guidelines encourage a kind of "cookbook medicine," where all patients are treated the same way. That isn't true. A good chef knows that a cookbook provides an important starting point, but that there are

usually several equally good options, depending on what ingredients are available and the desired outcomes. Similarly, the interpretation of a test and subsequent management decisions depend on the probability of disease. One example is the difference between a low-prevalence primary care or screening population and a high-prevalence referral or diseased population. For example, an abnormal CA-125 test followed by ultrasonography if the results are abnormal is 57 percent sensitive and 99 percent specific for ovarian cancer (positive LR = 57; negative LR = 0.43).[12] Therefore, this test is better at ruling in ovarian cancer when positive than at ruling it out when negative. But the prevalence of disease is critical in determining whether to use the test in practice. In the general population, in which the prevalence of ovarian cancer is only 0.04 percent,[13] the probability that a woman with an abnormal CA-125 test plus abnormal ultrasonography has ovarian cancer is only 2.2 percent. Using this test widely for screening would result in psychological harm and overuse of invasive testing and laparoscopy.[14] On the other hand, the test may be a sensible option in a high-prevalence population, such as women with a *BRCA1* or *BRCA2* mutation.

Combining Clinical Findings

Clinical decision rules combine findings from several elements of the history and physical examination, and sometimes a laboratory test, to help us make better diagnoses and prognoses. Well-known examples include the strep score[15] and Ottawa Ankle Rules,[16] but hundreds of others have been published, and many have been prospectively validated—something to look for before using them in the care of your own patients. PubMed's Clinical Queries Web site (http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml) and the Point-of-Care Guides featured in *American Family Physician* can be used to find clinical decision rules.

Most clinical decision rules place a patient in a risk group. This information can be used to guide further clinical decision-making. In general, when subsequent diagnostic tests are negative in a low-risk patient or positive in a high-risk patient, no further testing is necessary. Discordant results between the clinical rule and subsequent testing should prompt further evaluation. Remember, these are clinical decision-support tools, not clinical decision-replacement tools. They can improve our decision-making, but only if used wisely.

The Author

MARK H. EBELL, MD, MS, is a faculty member at the University of Georgia, Athens. He also is deputy editor for evidence-based medicine for *American Family Physician*. Dr. Ebell received his medical degree from the University of Michigan Medical School, Ann Arbor, where he also completed a family practice residency and received a master's degree in clinical research design and statistical analysis.

Address correspondence to Mark H. Ebell, MD, MS, 150 Yonah Ave., Athens, GA 30606 (e-mail: ebell@uga.edu). Reprints are not available from the author.

REFERENCES

1. Moe GW, Howlett J, Januzzi JL, Zowall H, for the Canadian Multicenter Improved Management of Patients With Congestive Heart Failure (IMPROVE-CHF) Study Investigators. N-terminal pro-B-type natriuretic peptide testing improves the management of patients with suspected acute heart failure: primary results of the Canadian prospective randomized multicenter IMPROVE-CHF study. *Circulation*. 2007;115(24):3103-3110.

2. Trowbridge RL, Rutkowski NK, Shojania KG. Does this patient have acute cholecystitis? *JAMA*. 2003;289(1):80-86.

3. Adepoju LJ, Chun J, El-Tamer M, Ditkoff BA, Schnabel F, Joseph KA. The value of clinical characteristics and breast-imaging studies in predicting a histopathologic diagnosis of cancer or high-risk lesion in patients with spontaneous nipple discharge. *Am J Surg*. 2005;190(4):644-646.

4. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron deficiency anemia: an overview [published correction appears in *J Gen Intern Med*. 1992;7(4):423]. *J Gen Intern Med*. 1992;7(2):145-153.

5. Katz JN, Dalgas M, Stucki G, et al. Degenerative lumbar spinal stenosis. Diagnostic value of the history and physical examination. *Arthritis Rheum*. 1995;38(9):1236-1241.

6. Smetana GW. The diagnostic value of historical features in primary headache syndromes: a comprehensive review. *Arch Intern Med*. 2000;160(18):2729-2737.

7. Goff BA, Mandel LS, Melancon CH, Muntz HG. Frequency of symptoms of ovarian cancer in women presenting to primary care clinics. *JAMA*. 2004;291(22):2705-2712.

8. Stoffers HE, Kester AD, Kaiser V, Rinkens PE, Knottnerus JA. Diagnostic value of signs and symptoms associated with peripheral arterial occlusive disease seen in general practice: a multivariable approach. *Med Decis Making*. 1997;17(1):61-70.

9. The PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *JAMA*. 1990;263(20):2753-2759.

10. Karlsson B, Granberg S, Wikland M, et al. Transvaginal ultrasonography of the endometrium in women with postmenopausal bleeding—a Nordic multicenter study. *Am J Obstet Gynecol*. 1995;172(5):1488-1494.

11. Brown MD, Rowe BH, Reeves MJ, Bermingham JM, Goldhaber SZ. The accuracy of the enzyme-linked immunosorbent assay D-dimer test in the diagnosis of pulmonary embolism: a meta-analysis. *Ann Emerg Med*. 2002;40(2):133-144.

12. Jacobs I, Davies AP, Bridges J, et al. Prevalence screening for ovarian cancer in postmenopausal women by CA 125 measurement and ultrasonography. *BMJ*. 1993;306(6884):1030-1034.

13. National Institutes of Health Consensus Development Conference Statement. Ovarian cancer: screening, treatment, and follow-up. *Gynecol Oncol*. 1994;55(3 pt 2):S4-S14.

14. Schapira MM, Matchar DB, Young MJ. The effectiveness of ovarian cancer screening. A decision analysis model. *Ann Intern Med*. 1993;118(11):838-843.

15. McIsaac WJ, Goel V, To T, Low DE. The validity of a sore throat score in family practice. *CMAJ*. 2000;163(7):811-815.

16. Stiell IG, Greenberg GH, McKnight RD, et al. Decision rules for the use of radiography in acute ankle injuries. Refinement and prospective validation. *JAMA*. 1993;269(9):1127-1132.