Test Properties 1:
Sensitivity, Specificity, and Predictive Values

Stuart Spitalnic, MD

In emergency department patients with chest pain who are at low risk for coronary disease, an ankle radiograph has a negative predictive value of 96% for coronary disease.

If the "dice test" is considered positive for pneumonia if a 2 or greater is rolled, the test's sensitivity is 83%.

I t would be ridiculous to use tests like these clinically, and the reason is clear: although the sensitivity and predictive values are "high," these tests do not add to our confidence regarding the presence or absence of disease in any given patient. I would suggest to you that, although never as irrelevant as the fake tests above, there are many real-life examples of tests that tout "high" values for a particular parameter, and seem as though they should be able to add important discriminatory information, yet in reality, the test's application adds little additional clinical information. Your job is to figure out which tests and testing strategies aid your clinical decision-making and which provide only a false sense of confidence. This is the first of a 2-part series reviewing test properties. Part 1 reviews sensitivity, specificity, and positive and negative predictive values. Part 2 will build on these concepts and discuss likelihood ratios, Bayes' theorem, and receiver-operating characteristics. Both articles include a few sample problems for practice in applying the concepts discussed.

Although there is more to interpreting the value of tests than knowing the definitions of the various tests properties, it is, nonetheless, essential to know their meanings. You will be unable to appreciate the nuances of spectrum bias and Bayes' theorem if you are struggling to remember whether the pneumonic is SPin and SNout or Sin and SPout. (It is the former—specificity is related to a test's ability to rule *in* a condition, sensitivity to a test's ability to rule *out* a condition. That being said, forget the mnemonic and learn the definitions.)

DEFINITIONS

Sensitivity: the probability that a test will be positive given a patient with the condition.

Specificity: the probability that a test will be negative given a patient without a condition.

Positive predictive value (PPV): the probability that a patient will have a condition given a positive test result.

Negative predictive value (NPV): the probability that a patient will not have a condition given a negative test result.

Traditionally, the definitions of the test properties are illustrated graphically in a 2×2 table, where one side represents the test result and the other side represents the true condition of the subject being tested (**Figure 1**). Looking at Figure 1, the *sensitivity* (the probability of testing positive given the presence of a condition) would be $A/(A + C)$; the *specificity* (the probability of testing negative in the absence of a condition) would be $D/(B + D)$; the *positive predictive value* (the probability of having a condition given a positive test) is $A/(A + B)$; and the *negative predictive value* (the probability of not having a condition given a negative test) is $D/(C + D)$. The *prevalence* (the proportion of the population that has a condition) is $(A + C)/(A + B + C + D)$.

BINARY QUESTIONS, CONTINUOUS ANSWERS

The serum human chorionic gonadotropin (hCG) level can range from zero to several hundred thousand; we want to know if a patient is pregnant. Blood glucose can range from 0 to more than 1000; we want to know if a pregnant woman has gestational diabetes. A leukocyte count can range from neutropenic to leukemic; we want to know whether a patient with abdominal pain has appendicitis. Most clinical questions are of the yes/no variety; most test results are on a continuum. How is this resolved?

Figure 2 presents results of a hypothetical test. The dotted line represents the range of results obtained in

Dr. Spitalnic is an assistant residency director, Brown University Emergency Medicine Residency Program, and an assistant professor of medicine, Brown University, Providence, RI.

	Has condition	**Does not have condition**	
Test positive	A	B	Total positive tests (A + B)
Test negative	C	D	Total negative tests (C + D)
	Number in sample with condition (A + C)	Number in sample without condition (B + D)	Total number of subjects (A + B + C + D)

Figure 1. The generic 2 × 2 table.

Figure 2. Hypothetical test results for patients with and without disease.

those without disease; the dashed line represents the range obtained in those with disease. Clearly, the average result in those with disease is higher than the average result in those without disease. When we are confronted with an individual patient, we want to know if the test can help determine whether the patient has or does not have a disease.

Most tests have a reported range of normal, reflecting the values that are likely to be obtained in patients who are normal (ie, without disease), the implication being that those outside the range are abnormal (ie, test positive for a disease). In Figure 2, if the closed arrows are considered the range of normal, those with test results to the right of the right-hand closed arrow would be considered to have a positive test. Notice how by using the right-hand closed arrow as the cut-off, we fail to label as positive those patients with disease whose results lie in the area of overlap. If we consider the test to be positive if the result is to the right of the left-hand open arrow, we would pick up all with disease, but misclassify as positive those who are normal whose test results are in the overlap area. The cut-off value can be arbitrarily decreased, resulting in more people with the condition being correctly identified at the expense of more people without the condition incorrectly testing positive. (That is, more true positives but also more false positives.) Raising the cut-off value will decrease the number of false positives but will result in more false negatives. We will revisit this concept when we discuss receiver-operating characteristics in the next article in this series, but the essential concepts for now are: (1) Diagnostic tests with continuous results require a cut-off value, above which the results of the test are considered positive; and (2) Overlap between the range of test results obtained for normal and abnormal patients
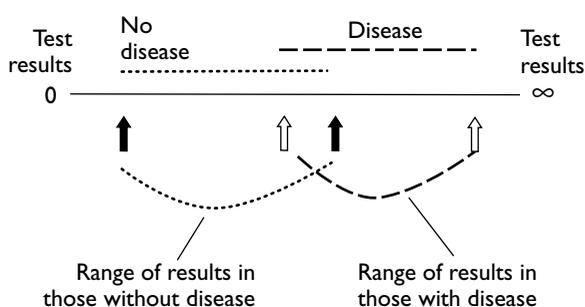
will, depending on the cut-off value selected, result in some healthy patients falsely testing positive and some diseased patients falsely testing negative.

DETERMINING TEST PROPERTIES: REFERENCE STANDARDS

Remember that a test's sensitivity is the probability it will be positive in those with disease; its specificity is the probability it will be negative in the absence of disease. This implies that in order to determine a test's sensitivity and specificity, the test must be applied to a group of patients in whom the diagnosis is known. To properly determine a test's sensitivity and specificity, it must be compared to a reference standard (often called a *gold standard*). For example, if you were trying to determine the test properties for D-dimer in the diagnosis of lower extremity deep venous thrombosis (DVT), ideally all patients would undergo a test that defines the illness (ie, venography), and simultaneously have their D-dimer levels measured. Patients could then be classified based on the reference test as positive or negative for DVT. D-dimer test results can be classified as above a diagnostic cut-off (positive) or below (negative). Suppose that, in a hypothetical sample of 500 patients, 100 had DVT proven by venography. Of the 100 with DVT, 90 have a positive D-dimer test, and of the 400 without DVT, 160 have a positive D-dimer test. These results are shown in **Figure 3**.

Now, given these data, what is the sensitivity of the D-dimer test for DVT? Sensitivity is the probability of a positive test given a patient with the disease (in Figure 1, $A/[A + C]$); therefore, the proportion is 90/100 (90 positive tests in the 100 that have disease) or 0.9, or, as a percent, 90%. Specificity is the probability of a negative test given the absence of disease $(D/[B + D])$; therefore, that proportion is 240/400 or 0.6 (60%).

The essential concept here is that a test's sensitivity

	Venography positive (+ DVT)	Venography negative (– DVT)	Total
D-dimer positive (test +)	90	160	250
D-dimer negative (test –)	10	240	250
Total	100	400	500

Figure 3. Deep venous thrombosis (DVT) data—scenario 1.

	Venography positive (+ DVT)	Venography negative (– DVT)	Total
D-dimer positive (test +)	225	100	325
D-dimer negative (test –)	25	150	175
Total	250	250	500

Figure 4. Deep venous thrombosis (DVT) data—scenario 2.

and specificity are determined by comparing the test's results to those of a reference standard. The expectation is that the test will behave the same way when applied to a similar group of patients. If, in the hypothetical study presented above, the D-dimer test was used in emergency department patients with leg pain and swelling and no risk factors for DVT, you could expect that if you applied this test to your emergency department patients who also had leg pain and swelling and no risk factors for DVT, the test would be positive in 90% of those with DVT but also would be falsely positive in 40% of those without DVT. (The false-positive proportion is equal to 1 – specificity.)

PREDICTIVE VALUES

When evaluating the power of a diagnostic test to discriminate between those with and without a condition, we are interested in the test's sensitivity and specificity. When we are faced with a patient and a test result and need to determine the likelihood that a patient has a condition (or does not have it), we are interested in the test's predictive value. When we say a test has an 80% PPV, it means that 80% of those with a positive test will actually have the condition. How are predictive values calculated?

Using the D-dimer data from Figure 3, what is the predictive value of a positive D-dimer test? The probability of those with a positive test having a DVT is 90/250 (A/[A + B]), or 36%. What about the NPV? Of the 250 with a negative test, only 10 had DVT, so the probability of not having a DVT given a negative test is 240/250 (D/[C + D]) or 96%.

What if, instead of 20% (100/500) of the population having DVT proven by venography, 50% (250/500) did? Presuming the same D-dimer test was performed

	Venography positive (+ DVT)	Venography negative (– DVT)	Total
D-dimer positive (test +)	18	192	210
D-dimer negative (test –)	2	288	290
Total	20	480	500

Figure 5. Deep venous thrombosis (DVT) data—scenario 3.

and that it behaved the same way (ie, had the same specificity and sensitivity—60% and 90%, respectively), the test results in the new population are shown in **Figure 4**. In this scenario, the PPV is 225/325, or 69%; the NPV is 150/175, or 86%. **Figure 5** shows the test results if the prevalence of DVT had been 4% instead of 20% (again, given the same specificity and sensitivity). Now, when the prevalence is 4% but the sensitivity and specificity remain the same, the PPV falls to 18/210, or 9%, but the NPV climbs to 99%.

The essential concept here is that a test's sensitivity and specificity are properties of the test and should be consistent when the test is used in similar patients in similar settings. Predictive values, although related to a test's sensitivity and specificity, will vary with the prevalence of the condition being tested for.

Let us revisit the statements made at the opening of this article regarding the ridiculous tests for coronary

artery disease and pneumonia and see why, although they are ridiculous, they are true:

The first statement suggested that an ankle radiograph has a NPV for coronary disease of 96%. The assumption (unstated) was that chest pain patients would have negative ankle radiographs. Also unstated, but approximately true, is that low-probability chest pain patients have a 4% incidence of coronary disease. So, if all ankle films are negative in a population of chest pain patients, the probability of not having disease given a negative ankle film is 96%. The test (as one would expect) adds nothing to your knowledge of the patient, and in fact has a sensitivity of zero, but nonetheless, an article could headline a 96% NPV for the ankle film test.

The second example suggested that a dice roll of 2 or greater was 83% sensitive for pneumonia. If sensitivity is the probability of a positive test given the presence of disease, the probability of rolling a 2 or greater with a patient who has pneumonia *is* 83%. (It also is 83% in those without pneumonia, but that is not the question.) So, the sensitivity of the "dice-test" is indeed 83%.

The next article in this series will expand on the concept of test properties, introducing likelihood ratios, Bayes' formula, and receiver-operating characteristic curves. The following section presents a few problems relating to what has been discussed in this article.

PRACTICE QUESTIONS

1. The figure below presents leukocyte counts (× 10³/mm³) for 20 children with fever aged 3 months to 1 year. The column of the left represents children with negative blood cultures, and the column on the right represents those with positive blood cultures:

(–) Blood Cultures	(+) Blood Cultures
7	11.1
7.2	14.6
7.2	15.5
8.6	18.3
9	26.7
9.1	
10.1	
10.2	
11	
11.1	
12.1	
12.2	
14.7	
15.1	
17	

- What is the prevalence of bacteremia?

- What are the sensitivity and specificity for bacteremia of a leukocyte count greater than 10? Greater than 12? Greater than 15?

2. A patient comes to your office frantic over the results of a home HIV test. The test touts 99% sensitivity and 99% specificity. On questioning, you determine that this patient is at low risk for HIV; given your assessment of his risk factors, you believe he comes from a population group that has a baseline prevalence of HIV of 1 in 100,000. He now presents to you with a positive result on his home HIV test. Given his baseline risk and the positive home test, what are the chances that this patient is actually HIV positive?

ANSWERS

1. *Prevalence* is the proportion of patients who have a condition at a particular time. In the sample of 20 patients, 5 had bacteremia. The prevalence is 5/20, or 25%. *Sensitivity* is the proportion of those with a positive test given that they have the condition being tested for. When calculating the sensitivity, you only need be concerned with the 5 patients who have bacteremia. All patients with bacteremia had leukocyte counts greater than 10, so the sensitivity of a leukocyte count greater than 10 is 100%. When a leukocyte count of 12 is used as the cut-off, the sensitivity is 80% (4 positive out of 5). For a leukocyte count of 15, 3/5 are positive, for a sensitivity of 60%. *Specificity* is the proportion of those with a negative test given that they do not have the condition. Here, only those whose blood cultures are negative need be considered. When a leukocyte count of 10 is used as the cut-off for a positive test, of the 15 without bacteremia, 6 will correctly test negative, for a specificity of 6/15, or 40%. When 12 is used as the cut-off value, the specificity is 10/15, or 67%. When 15 is used, the specificity is 13/15, or 87%. Notice—and this will be discussed in future articles—that when you change the cut-off value for a positive test, the sensitivity and specificity move in opposite directions.

2. Question 2 can be summarized as follows: given a test with a sensitivity and specificity of 99% and a condition's prevalence of 1 in 100,000, what is the predictive value of a positive test? (The PPV is the proportion of patients who will have a condition, given that they have a positive test.) There are many ways to solve this problem; the most illustrative is to

construct a 2 × 2 table based on a hypothetical population. Although any numbers can be used at the outset, it is often convenient to consider a population that is 100 times greater than the prevalence's denominator; in this case, that would be 10 million. In a population of 10 million with a prevalence of 1 in 100,000, you would expect 100 people to actually be HIV-positive and 9,999,900 to be HIV-negative. A sensitivity of 99% means that of the 100 with HIV, 99 will test positive and 1 will test negative. A specificity of 99% means that of the 9,999,900 without HIV, 9,899,901 will test negative and 99,999 will test positive. The 2 × 2 table would then look like this:

	HIV +	HIV –	Total
Test +	99	99,999	100,098
Test –	1	9,899,901	9,899,902
Total	100	9,999,900	10,000,000

The PPV is the proportion of those who have the condition given a positive test, in this case, 99/100,098, or 0.001 (0.1%). Notice that even though the sensitivity and specificity are 99%, with the low prevalence, approximately 1000 false-positive tests occur for every true positive. On your own, calculate the PPV for the same test if the prevalence were 1 in 1000. What if the prevalence were 1 in 100? (The answers are 9% and 50%, respectively.)   **HP**

EDITOR'S NOTE

The first article in the *Primer in Literature Interpretation* series, "Clinician's Probability Primer," appeared in the February 2003 issue of *Hospital Physician* and can be downloaded at our web site (www.turner-white.com).