8 • THE SENSITIVITY, SPECIFICITY, AND PREDICTIVE VALUE OF DIAGNOSTIC TESTS

RICHARD C. REBA • JOEL C. KLEINMAN

In attempting to determine the utility of a particular test procedure from a review of the literature, it appears that one group is frequently unable to reproduce the test results reported by others. The reporting of such opposing conclusions might lead one to infer that the test procedure is of no value because the test is unreliable and cannot be reproduced. However, the apparent difference in conclusions is often a result of a lack of clear understanding of the terms used to evaluate test performance and the variables that influence these terms.

In the evaluation of the usefulness of a particular test, it is important to differentiate the *sensitivity* and the *specificity* (and the related false-positive and false-negative rates) from the *conditional probability of a disease being present, given a positive test* (or of "no disease," given a negative test). The relationships among these various probabilities will be described in this chapter. The major point will be that evaluation of diagnostic tests depends upon the prevalence of disease in the population to which the test is applied.

Definitions

Suppose a test is applied to a population of N people. The test results can be described as in Table 8-1, assuming the true state of each person is known. The symbols in the diagonal cells refer to true test results (TP = true-positive and TN = true-negative numbers), and the off-diagonal cells are false test results (FN = false-negative and FP = false-positive numbers). A perfect test would have no persons in the off-diagonal cells.

The crucial measures of the power of a test to distinguish diseased from healthy persons are *sensitivity*, the probability of being able to identify correctly those who DO have a disease, and *specificity*, the probability of being able to identify correctly those who DO NOT have the disease. In Table 8-1, these parameters are measured as follows:

$$\text{Sensitivity (SN)} = TP/N_P = \text{probability of positive test, given patient has disease}$$

$$\text{Specificity (SP)} = TN/N_A = \text{probability of negative test, given patient does not have disease}$$

In a perfect test, both sensitivity and specificity are equal to 1.

Table 8-1. Measurement of test effectiveness

Test result	Disease present	Disease absent	Totals
+	TP	FP	N+
−	FN	TN	N−
Totals	N_P	N_A	N

Supported in part by USPHS Grants GM 20543 and CA 16284

The *rate* of false-negatives (RFN) is FN/N_P, which is 1 minus the sensitivity.

$$RFN = FN/N_P = 1 - SN$$

Similarly, the *rate* of false-positives (RFP) is FP/N_A, which is 1 minus the specificity.

$$RFP = FP/N_A = 1 - SP$$

Since false-negatives and false-positives also refer to numbers, the terms sensitivity and specificity should be used to avoid confusion (Fig. 8-1).

The terms *sensitivity* and *specificity* refer to conditional probability statements defining the ability of a test to accurately identify the presence or absence of a disease in a population of tested individuals. Therefore, if a test is determined to have 90% sensitivity, one can expect to find abnormal values in 90% of all those persons tested who have the disease; 10% of diseased subjects will have a false-negative test result. Analogously, if a test has been determined to have 95% specificity, then one can expect a normal reading in 95% of all healthy subjects; 5% of the healthy subjects will have a false-positive test result.

It is important to note that these measures of the test's effectiveness use in their denominators the numbers of persons with and without the disease, not the numbers of patients with positive and negative tests (N+ and N−). Therefore, calculation of these parameters requires the ability to diagnose *definitively* "disease" and "non-disease" by other independent means. Thus, when a new test for a disease is being evaluated, it is necessary to perform the test in two selected groups of subjects. The true sensitivity is determined in a group of subjects with unqualified diagnosis of the disease by other criteria. The true specificity of a test must be determined in a *second* group of selected and controlled patients in whom the absence of the disease, asymptomatic or early disease, or factors that predispose or result in a higher risk of the disease have been excluded by other independent and accurate tests or procedures.

However, what sometimes happens in

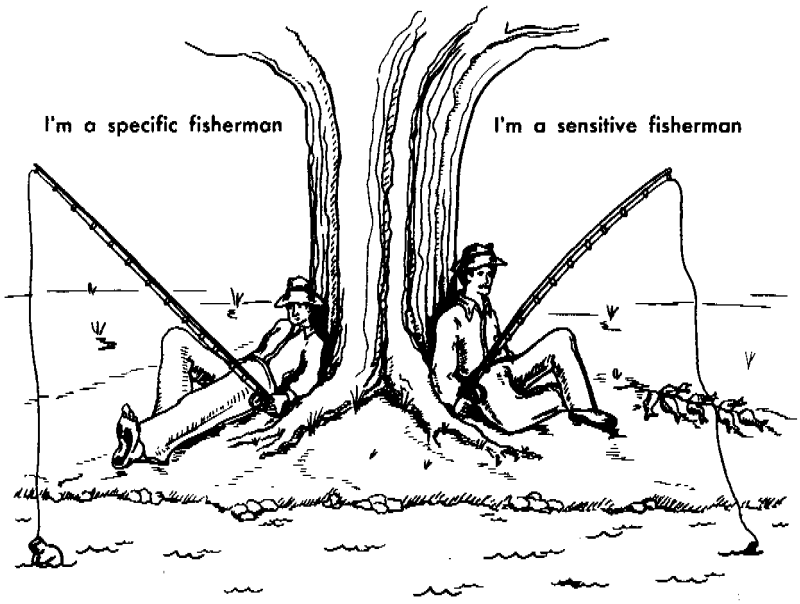

Fig. 8-1. The specific fisherman catches *only* big fish; the sensitive fisherman catches all big fish and a lot of little ones, too.

practice is that the test is evaluated in one population, namely, a group of subjects with the disease. This procedure enables the investigators to determine whether their proposed test can correctly identify patients known to have the disease. It is important to determine sensitivity first since if the new test cannot identify disease correctly, there is no need to continue with it. However, once this is accomplished, usually by investigators interested in a particular disease or in a referral center (in either instance a population with a large proportion of diseased subjects in the test group), it may be difficult for these same workers to have access to a healthy population having a low disease prevalence. Therefore, information regarding the ability of this test to exclude patients who do not have the disease, that is, specificity, will be unavailable. This will usually result in disappointment when the test is used to screen large unselected populations with a low prevalence of the disease.

The relationship between sensitivity and specificity is best illustrated by considering a test that results in a continuous measurement, for example, blood glucose. Fig. 8-2 shows the distributions of the measurement in diseased and healthy populations. Any result greater than x will be considered positive. Since the populations overlap, any cutoff point will involve error. As the cutoff point increases, the number of false-negatives increases, and the number of false-positives decreases. Thus, the test's sensitivity will decrease while the specificity increases. The appropriate cutoff point will depend upon the relative costs and benefits of different levels of diagnostic accuracy, therapeutic impact, patient outcome, and the population to which the test will be applied.

This last point leads to the next two parameters, which are needed to evaluate screening tests. In the population to be tested, what is the probability that a patient with a positive test has the disease and

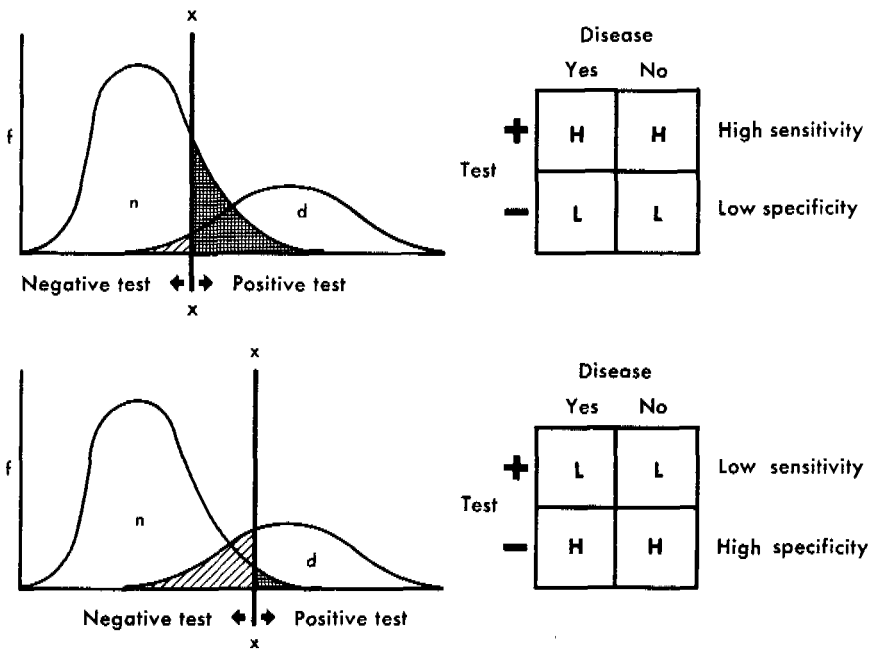

Fig. 8-2. Overlapping frequency distributions of test results obtained on subjects without disease (normal, or n) and subjects with disease (d). Because of the overlapping frequency distributions, the sensitivity and specificity are determined by the decision boundary line, x.

that a patient with a negative test does not have the disease? These probabilities are a function of three values: (1) sensitivity of the test, (2) specificity of the test, and (3) prevalence of the disease in the population. They can be derived directly from Table 8-1 as TP/N+ and TN/N−. However, when this is done, the explicit role of prevalence is not evident. This is precisely how the confusion mentioned earlier can arise. By evaluating the test on different populations with different disease prevalences, these probabilities will vary.

To illustrate this point the two probabilities can be derived from Bayes' theorem.[1] The first is called the *predictive value of a positive test result:*[7]

$$PV+ = \frac{SN \times DP}{(SN \times DP) + (1 - SP) \times (1 - DP)} =$$

probability of disease, given positive test result

(DP = disease prevalence. PV+ measures the degree to which a positive test confirms the diagnosis.)

The second is called the *predictive value of a negative test result:*

$$PV- = \frac{SP \times (1 - DP)}{(1 - SN) \times DP + SP \times (1 - DP)} =$$

probability of no disease, given negative test result

(PV− measures the degree to which a negative test excludes the diagnosis.)

For example, if a liver scan (sensitivity = 71% and specificity = 95%)[5] is performed on a patient from a family physician's practice in which a reasonable estimate of the prevalence of hepatic malignancy is 1%, an abnormal liver scan results in a probability of only 12.5% that the patient has hepatic metastases (PV+). In this case, PV− is 99.7%. However, if a liver scan is performed and the same interpretation criteria are used on a patient from an oncologist's practice or from a hospital inpatient population in which the prevalence of malignancy is 25%, a positive liver scan results in an 82.5% probability that the patient truly has hepatic metastases. For this population, PV− is 90.8%. In the first case, if the probability is increased from 1% to 13% that malignancy is present, the family physician probably would not recommend liver biopsy or peritoneoscopy to confirm the diagnosis. However, in the second instance, when the probability is increased from 25% to 83%, the recommendation for peritoneoscopy and biopsy is on much firmer ground. Thus, although the test is precisely the same in each situation, its operating characteristics are very different as a result of the different disease prevalences.

Table 8-2 shows the predictive value of positive and negative tests for a range of values of sensitivity, specificity, and disease prevalence. First, note the effect of prevalence: even with 99% sensitivity and specificity, the predictive value of a positive test can be as low as 9% for a very rare disease (0.1% prevalence) and rise to 84% for a disease with 5% prevalence or to over 99% for a disease with 50% prevalence. Prevalence does not have such a marked effect on PV− unless the prevalence is very high or sensitivity and specificity are low.

However, just because a test is shown to have a relatively low PV+, it should not be concluded that the test is of little value. Katz[6] has implied that a test that results in a PV+ of 50% should not be performed.[4] This is an incorrect interpretation of predictive value since it ignores the facts that prior to the test the probability that the patient had the disease may have been much lower than 50% (10% in Katz's example) and that PV− may be high enough to rule out the disease. For example, in a population of 1,000 with a disease prevalence of 10%, a test with 90% sensitivity and specificity will result in 100 positive results, of which 50 patients will have the disease. Subsequent tests on the 100 patients will be necessary, but the size of the population will have been substantially reduced. Of course, this was done at the expense of missing 50 patients with disease. Only a test with greater specificity would significantly reduce the number of false-negative tests (and thus increase PV+), as can be seen in Table 8-2. The best way to in-

Table 8-2. Predictive value of positive and negative test results

Specificity (%)	PV+ (%) Sensitivity (%)								PV− (%) Sensitivity (%)							
	10	30	50	70	90	95	98	99	10	30	50	70	90	95	98	99
	Prevalence = 0.1%								*Prevalence* = 0.1%							
10	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	99.1	99.3	99.5	99.7	99.9	100.0	100.0	100.0
30	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	99.7	99.8	99.8	99.9	100.0	100.0	100.0	100.0
50	0.0	0.1	0.1	0.1	0.2	0.2	0.2	0.2	99.8	99.9	99.9	99.9	100.0	100.0	100.0	100.0
70	0.0	0.1	0.2	0.2	0.3	0.3	0.3	0.3	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0
90	0.1	0.3	0.5	0.7	0.9	0.9	1.0	1.0	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0
95	0.2	0.6	1.0	1.4	1.8	1.9	1.9	1.9	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0
98	0.5	1.5	2.4	3.4	4.3	4.5	4.7	4.7	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0
99	1.0	2.9	4.8	6.5	8.3	8.7	8.9	9.0	99.9	99.9	99.9	100.0	100.0	100.0	100.0	100.0
	Prevalence = 0.5%								*Prevalence* = 0.5%							
10	0.1	0.2	0.3	0.4	0.5	0.5	0.5	0.5	95.7	96.6	97.5	98.5	99.5	99.7	99.9	99.9
30	0.1	0.2	0.4	0.5	0.6	0.7	0.7	0.7	98.5	98.8	99.2	99.5	99.8	99.9	100.0	100.0
50	0.1	0.3	0.5	0.7	0.9	0.9	1.0	1.0	99.1	99.3	99.5	99.7	99.9	99.9	100.0	100.0
70	0.2	0.5	0.8	1.2	1.5	1.6	1.6	1.6	99.4	99.5	99.6	99.8	99.9	100.0	100.0	100.0
90	0.5	1.5	2.5	3.4	4.3	4.6	4.7	4.7	99.5	99.6	99.7	99.8	99.9	100.0	100.0	100.0
95	1.0	2.9	4.8	6.6	8.3	8.7	9.0	9.0	99.5	99.6	99.7	99.8	99.9	100.0	100.0	100.0
98	2.5	7.0	11.2	15.0	18.4	19.3	19.8	19.9	99.5	99.6	99.7	99.8	99.9	100.0	100.0	100.0
99	4.8	13.1	20.1	26.0	31.1	32.3	33.0	33.2	99.5	99.6	99.7	99.8	99.9	100.0	100.0	100.0
	Prevalence = 1.0%								*Prevalence* = 1.0%							
10	0.1	0.3	0.6	0.8	1.0	1.1	1.1	1.1	91.7	93.4	95.2	97.1	99.0	99.5	99.8	99.9
30	0.1	0.4	0.7	1.0	1.3	1.4	1.4	1.4	97.1	97.7	98.3	99.0	99.7	99.8	99.9	100.0
50	0.2	0.6	1.0	1.4	1.8	1.9	1.9	2.0	98.2	98.6	99.0	99.4	99.8	99.9	100.0	100.0
70	0.3	1.0	1.7	2.3	2.9	3.1	3.2	3.2	98.7	99.0	99.3	99.6	99.9	99.9	100.0	100.0
90	1.0	2.0	4.8	6.6	8.3	8.8	9.0	9.1	99.0	99.2	99.4	99.7	99.9	99.9	100.0	100.0
95	2.0	5.7	9.2	12.4	15.4	16.1	16.5	16.7	99.1	99.3	99.5	99.7	99.9	99.9	100.0	100.0
98	4.8	13.2	20.2	26.1	31.3	32.4	33.1	33.3	99.1	99.3	99.5	99.7	99.9	99.9	100.0	100.0
99	9.2	23.3	33.6	41.4	47.6	49.0	49.7	50.0	99.1	99.3	99.5	99.7	99.9	99.9	100.0	100.0
	Prevalence = 2.0%								*Prevalence* = 2.0%							
10	0.2	0.7	1.1	1.6	2.0	2.1	2.2	2.2	84.5	87.5	90.7	94.2	98.0	99.0	99.6	99.8
30	0.3	0.9	1.4	2.0	2.6	2.7	2.8	2.8	94.2	95.5	96.7	98.0	99.3	99.7	99.9	99.9
50	0.4	1.2	2.0	2.8	3.5	3.7	3.8	3.9	96.5	97.2	98.0	98.8	99.6	99.8	99.9	100.0
70	0.7	2.0	3.3	4.5	5.8	6.1	6.3	6.3	97.4	98.0	98.6	99.1	99.7	99.9	99.9	100.0

Continued.

Table 8-2. Predictive value of positive and negative test results—cont'd

Specificity (%)	PV+ (%) Sensitivity (%)								Specificity (%)	PV− (%) Sensitivity (%)							
	10	**30**	**50**	**70**	**90**	**95**	**98**	**99**		**10**	**30**	**50**	**70**	**90**	**95**	**98**	**99**
	Prevalence = 2.0%—cont'd									*Prevalence = 2.0%—cont'd*							
90	2.0	5.8	9.3	12.5	15.5	16.2	16.7	16.8	**90**	98.0	98.4	98.9	99.3	99.8	99.9	100.0	100.0
95	3.9	10.9	16.9	22.2	26.9	27.9	28.6	28.8	**95**	98.1	98.5	98.9	99.4	99.8	99.9	100.0	100.0
98	9.3	23.4	33.8	41.7	47.9	49.2	50.0	50.3	**98**	98.2	98.6	99.0	99.4	99.8	99.9	100.0	100.0
99	16.9	38.0	50.5	58.8	64.7	66.0	66.7	66.9	**99**	98.2	98.6	99.0	99.4	99.8	99.9	100.0	100.0
	Prevalence = 3.0%									*Prevalence = 3.0%*							
10	0.3	1.0	1.7	2.3	3.0	3.2	3.3	3.3	**10**	78.2	82.2	86.6	91.5	97.0	98.5	99.4	99.7
30	0.4	1.3	2.2	3.0	3.8	4.0	4.2	4.2	**30**	91.5	93.3	95.1	97.0	99.0	99.5	99.8	99.9
50	0.6	1.8	3.0	4.2	5.3	5.6	5.7	5.8	**50**	94.7	95.8	97.0	98.2	99.4	99.7	99.9	99.9
70	1.0	3.0	4.9	6.7	8.5	8.9	9.2	9.3	**70**	96.2	97.0	97.8	98.7	99.6	99.8	99.9	100.0
90	3.0	8.5	13.4	17.8	21.8	22.7	23.3	23.4	**90**	97.0	97.7	98.3	99.0	99.7	99.8	99.9	100.0
95	5.8	15.7	23.6	30.2	35.8	37.0	37.7	38.0	**95**	97.2	97.8	98.4	99.0	99.7	99.8	99.9	100.0
98	13.4	31.7	43.6	52.0	58.2	59.5	60.2	60.5	**98**	97.2	97.8	98.4	99.1	99.7	99.8	99.9	100.0
99	23.6	48.1	60.7	68.4	73.6	74.6	75.2	75.4	**99**	97.3	97.9	98.5	99.1	99.7	99.8	99.9	100.0
	Prevalence = 5.0%									*Prevalence = 5.0%*							
10	0.6	1.7	2.8	3.9	5.0	5.3	5.4	5.5	**10**	67.9	73.1	79.2	86.4	95.0	97.4	99.0	99.5
30	0.7	2.2	3.6	5.0	6.3	6.7	6.9	6.9	**30**	86.4	89.1	91.9	95.0	98.3	99.1	99.7	99.8
50	1.0	3.1	5.0	6.9	8.7	9.1	9.4	9.4	**50**	91.3	93.1	95.0	96.9	99.0	99.5	99.8	99.9
70	1.7	5.0	8.1	10.9	13.6	14.3	14.7	14.8	**70**	93.7	95.0	96.4	97.8	99.3	99.6	99.8	99.9
90	5.0	13.6	20.8	26.9	32.1	33.3	34.0	34.3	**90**	95.0	96.1	97.2	98.3	99.4	99.7	99.9	99.9
95	9.5	24.0	34.5	42.4	48.6	50.0	50.8	51.0	**95**	95.3	96.3	97.3	98.4	99.4	99.7	99.9	99.9
98	20.8	44.1	56.8	64.8	70.3	71.4	72.1	72.3	**98**	95.4	96.4	97.4	98.4	99.5	99.7	99.9	99.9
99	34.5	61.2	72.5	78.7	82.6	83.3	83.8	83.9	**99**	95.4	96.4	97.4	98.4	99.5	99.7	99.9	99.9
	Prevalence = 10.0%									*Prevalence = 10.0%*							
10	1.2	3.6	5.8	8.0	10.0	10.5	10.8	10.9	**10**	50.0	56.3	64.3	75.0	90.0	94.7	97.8	98.9
30	1.6	4.5	7.4	10.0	12.5	13.1	13.5	13.6	**30**	75.0	79.4	84.4	90.0	96.4	98.2	99.3	99.6
50	2.2	6.3	10.0	13.5	16.7	17.4	17.9	18.0	**50**	83.3	86.5	90.0	93.8	97.8	98.9	99.6	99.8
70	3.6	10.0	15.6	20.6	25.0	26.0	26.6	26.8	**70**	87.5	90.0	92.6	95.5	98.4	99.2	99.7	99.8
90	10.0	25.0	35.7	43.7	50.0	51.4	52.1	52.4	**90**	90.0	92.0	94.2	96.4	98.8	99.4	99.8	99.9
95	18.2	40.0	52.6	60.9	66.7	67.9	68.5	68.8	**95**	90.5	92.4	94.5	96.6	98.8	99.4	99.8	99.9
98	35.7	62.5	73.5	79.5	83.3	84.1	84.5	84.6	**98**	90.7	92.6	94.6	96.7	98.9	99.4	99.8	99.9
99	52.6	76.9	84.7	88.6	90.9	91.3	91.6	91.7	**99**	90.8	92.7	94.7	96.7	98.9	99.4	99.8	99.9

Specificity (%)	Sensitivity (%)							
				Prevalence = 30.0%				
10	4.5	12.5	19.2	25.0	30.0	31.1	31.8	32.0
30	5.8	15.5	23.4	30.0	35.5	36.8	37.5	37.7
50	7.9	20.5	30.0	37.5	43.5	44.9	45.7	45.9
70	12.5	30.0	41.7	50.0	56.3	57.6	58.3	58.6
90	30.0	56.3	68.2	75.0	79.4	80.3	80.8	80.9
95	46.2	72.0	81.1	85.7	88.5	89.1	89.4	89.5
98	68.2	86.5	91.5	93.8	95.1	95.3	95.5	95.5
99	81.1	92.8	95.5	96.8	97.5	97.6	97.7	97.7
				Prevalence = 50.0%				
10	10.0	25.0	35.7	43.8	50.0	51.4	52.1	52.4
30	12.5	30.0	41.7	50.0	56.3	57.6	58.3	58.6
50	16.7	37.5	50.0	58.3	64.3	65.5	66.2	66.4
70	25.0	50.0	62.5	70.0	75.0	76.0	76.6	76.7
90	50.0	75.0	83.3	87.5	90.0	90.5	90.7	90.8
95	66.7	85.7	90.9	93.3	94.7	95.0	95.1	95.2
98	83.3	93.8	96.2	97.2	97.8	97.9	98.0	98.0
99	90.9	96.8	98.0	98.6	98.9	99.0	99.0	99.0

Specificity (%)	Sensitivity (%)							
				Prevalence = 30.0%				
10	20.6	25.0	31.8	43.8	70.0	82.4	92.1	95.9
30	43.8	50.0	58.3	70.0	87.5	93.3	97.2	98.6
50	56.5	62.5	70.0	79.5	92.1	95.9	98.3	99.2
70	64.5	70.0	76.6	84.5	94.2	97.0	98.8	99.4
90	70.0	75.0	80.8	87.5	95.5	97.7	99.1	99.5
95	71.1	76.0	81.6	88.1	95.7	97.8	99.1	99.6
98	71.8	76.6	82.1	88.4	95.8	97.9	99.1	99.6
99	72.0	76.7	82.2	88.5	95.9	97.9	99.1	99.6
				Prevalence = 50.0%				
10	10.0	12.5	16.7	25.0	50.0	66.7	83.3	90.9
30	25.0	30.0	37.5	50.0	75.0	85.7	93.8	96.8
50	35.7	41.7	50.0	62.5	83.3	90.9	96.2	98.0
70	43.8	50.0	58.3	70.0	87.5	93.3	97.2	98.6
90	50.0	56.3	64.3	75.0	90.0	94.7	97.8	98.9
95	51.4	57.6	65.5	76.0	90.5	95.0	97.9	99.0
98	52.1	58.3	66.2	76.6	90.7	95.1	98.0	99.0
99	52.4	58.6	66.4	76.7	90.8	95.2	98.0	99.0

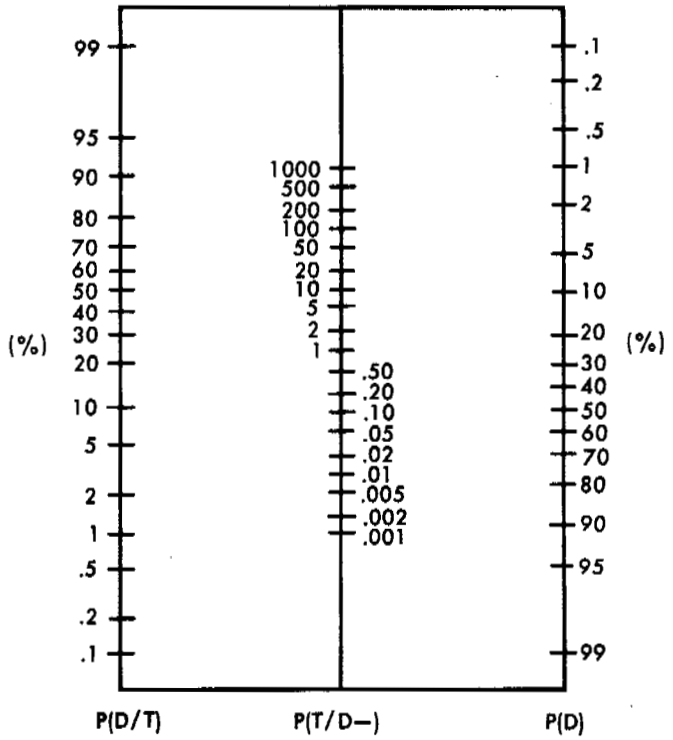

Fig. 8-3. Chart for computing predictive value.[3] P(D) denotes disease prevalence; P (T/D) denotes the probability of the test result (positive or negative) given that the disease is present; P (T/D−) denotes the probability of the test result given that the disease is absent; P (D/T) denotes the probability that the disease is present given the test result.

To find the predictive value of a positive test P(D+/T+), draw a line through the prevalence on the right and the center line, the ratio of P(T/D), sensitivity, to P(T/D−), one minus specificity, or the RFP. The point at which the line intersects the left axis is P(D/T+). To find the predictive value of a negative test P(D−/T−), draw a line through the prevalence on the right and the center line, the ratio of P(T/D), one minus the sensitivity, to P(T−/D−), specificity. The point at which the line intersects the left axis is P(D/T−), which is 100% minus the predictive value of a negative test.

crease PV+ is to increase the *specificity* of the test. Conversely, the most effective way to increase PV− is to increase *sensitivity*.

Fig. 8-3 provides an easy method to calculate PV+ and PV−, given the disease prevalence, and the test's sensitivity and specificity.[3]

Application of diagnostic tests in series

The previous definitions can be applied to a series of diagnostic tests to justify the intuitively apparent strategy of applying more than one diagnostic test to confirm a diagnosis. However, a crucial assumption in the justification is that the tests are *independent*. That is, a patient with the disease should have the same probability of being positive on the second test whether or not the results of the first test were positive. The same must hold for patients with no disease. This is true in only a limited sense for a second determination of the same diagnostic test (see discussion of reliability on p. 63).

Basically, the sequential application of two tests increases the disease prevalence in the group that requires the second test. For example, by applying a test with sensitivity and specificity of 90% to a population with disease prevalence of 10%, we obtain

a population of positives, 50% of whom have the disease (this is PV+). Thus, since the second test is applied only to the group of positives, the relevant prevalence is 50%. If this test has 95% sensitivity and specificity, PV+ is nearly 95%. PV− for the series of tests is nearly 98%. Applying the tests in the reverse order would result in the same values of PV+ and PV−. The major considerations in choosing which one to apply first are that the costs and discomfort involved should be less for the initial test since it is applied to the larger population.

The sensitivity of the series of two tests (the sequential test) is the product of the sensitivity of each test. The specificity of the sequential test is 1 minus the product of the rates of false-positives for each test, that is, $1 - (1 - SP_1)(1 - SP_2)$. Thus, the sequential test has a higher specificity than either test alone but a lower sensitivity. From Table 8-2, we noted earlier that increasing specificity increases the predictive value of a positive test more rapidly than increasing sensitivity. The sequential test takes advantage of this property by increasing specificity at the expense of a decrease in sensitivity. It can be shown that the sequential test has a higher PV+ than either individual test if for each individual test the probability of a positive result is greater for a diseased person than for a nondiseased person (i.e., if the sensitivity is greater than 1 minus the specificity). If this were not true for an individual test, that test would have a predictive value less than the disease prevalence and would be worse than no test at all. Thus, for all practical purposes, the sequential test has a higher PV+ than either individual test. PV+ for the sequential test can be determined from Table 8-2 or Fig. 8-2 by calculating the sensitivity and specificity of the sequential test (using the product rule given before) and applying it to the (original) population's disease prevalence.

The sequential test may have a smaller PV− than either individual test or it may have an intermediate value. In either case, PV− is usually close enough to 100% that the differences are negligible.

Using the previous example of a first test with sensitivity and specificity of 90% and a second test with 95% sensitivity and specificity, the sensitivity and specificity of the sequential test are as follows:

$$SN = .90 \times .95 = .855$$
$$SP = 1 - .10 \times .05 = .995$$

Applying the sequential test to a population with a disease prevalence of 10% results in a PV+ of 95% compared to a PV+ of 50% and 68% for the first and second tests respectively (applied alone). The predictive value of a negative test is 98.7% compared to 98.8% for the first test and 99.4% for the second.

The previous discussion has shown that one's effort is directed best to those patients having a positive test to continue the exclusion process by performing other independent tests. This intuitive operational strategy minimizes the overall probability of a false-positive error. Clinical skepticism usually demands a corroborative positive test that makes the risk of any single false-positive error small. Clinicians and biomedical scientists operate similarly in that more than one positive result is required before a significant conclusion is reached. It is rare to diagnose a disease, particularly a significant disease, on the basis of a single abnormal test result. The probability of two or three serial false-positive independent test results is exceedingly low. A good clinician traditionally is selective in ordering tests and believes in strong "indications" for ordering any test. In effect, what he is doing is setting the disease prevalence high before he orders a test and therefore significantly increases the predictive value of a positive test result.

The preceding discussion should be contrasted with a common situation in multiphasic screening where a battery of diagnostic tests are applied to determine whether a patient has one of a number of possible diseases. In the previous situation, sensitivity and specificity were evaluated under

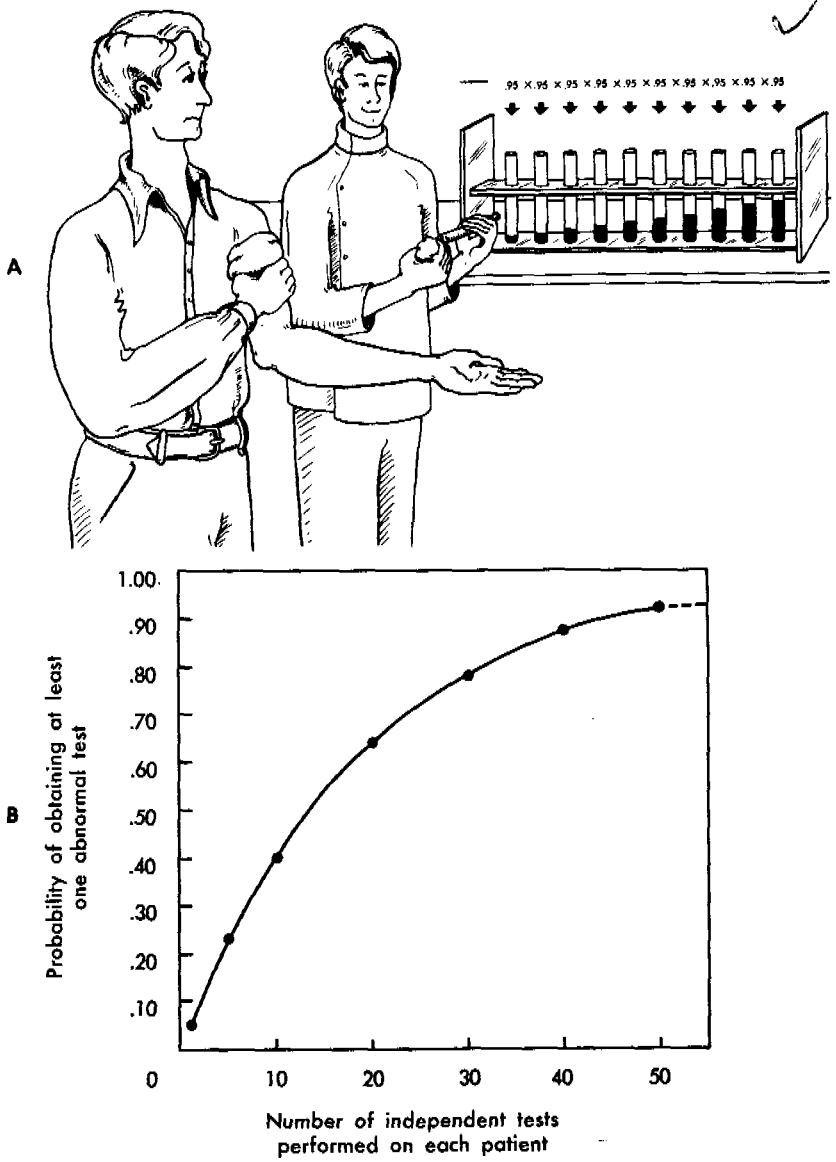

Fig. 8-4. A, "Keep testing, we'll find something abnormal." **B,** As the number of independent tests performed on each patient increases, so does the probability of obtaining *at least* one abnormal test.

the condition that an individual was considered positive only if two or more tests for the same disease were positive. In the multiphasic screening situation an individual may be followed if he is positive on *at least one* of the tests since the tests are usually concerned with different diseases. In this case the rate of false-positives—that is, the proportion of individuals having at least one positive test among those with no disease—may become quite large. For example, the normal range of many laboratory tests is set at ±2 standard deviations to give (approximately) 95% specificity. If five *independent* laboratory tests are applied, the proportion of normal individuals with at least one positive ("abnormal") test result is 23%. If ten tests are applied, there will

be a 40% rate of false-positives; if 20 tests are performed the rate of false-positives increases to 64%! The rate of false-negatives is more difficult to evaluate since its definition depends upon which diseases are actually present (Fig. 8-4).

Reliability

Thus far, this discussion of diagnostic tests has ignored the problem of reliability, or reproducibility, of test results. A full treatment of this topic is beyond the scope of this chapter, but a few comments are in order. First is the obvious point that an unreliable test (i.e., one in which repeated observations on identical samples give different results) can neither be very sensitive nor specific.

Related to this point is the fact that reliability needs to be evaluated as carefully as sensitivity and specificity, with full cognizance of the disease prevalence in the tested population (Fig. 8-1). If the measurement error (i.e., errors resulting from the instrument, the observer, or the laboratory procedure) is constant for all values of blood glucose, the reliability of the test will not depend upon the population tested. However, if the measurement error is greater at higher values of blood glucose (especially values near the cutoff point), the reliability will depend upon the population tested. If a predominantly normal population is tested, the reliability of the method may be overestimated.

A final point mentioned in the discussion of sequential tests is that repetition of the same diagnostic test is not equivalent to the application of two different tests. The former is a way to control reliability—for instance, by repeating an elevated blood glucose measurement to increase the probability that the patient's blood glucose is really abnormally high. The degree to which such a test reflects clinical diabetes (its sensitivity and specificity) cannot be measured this way. If the test is known to have limited reliability, it is of course wise to order a retest. However, this will not increase predictive value in the same way that two clinically different procedures do.

Conclusions

From this discussion, it should be apparent that in arriving at a diagnosis it is impossible to avoid risk. Indeed, all test interpretation, regardless of the experience of the observer, is associated with some error. In a most primitive form the error is of two kinds. Either the test states that a disease is present when it is not, or the observer interprets the test as negative and the sought-after disease is actually present.

Probabilities must be evaluated, but other types of risk or cost should also be considered. In addition to sensitivity, specificity, and both types of predictive value, the judgment as to whether a test should be used is related directly to how treatable the disease is and is related inversely to the risks involved in further diagnostic tests or subsequent therapy. Even if it is unlikely to be present, a disease entity should be included in the diagnostic workup if it is readily treatable and requires little or no risk in arriving at the diagnosis or in subsequent therapy. On the other hand, there is no urgency to confirm an untreatable disease, even if the likelihood of its being present is high, if the risk or cost of arriving at the diagnosis is great.

Since the various costs and the implication of a positive or negative test vary under different circumstances, it is worth noting that a determination of the specific worth of an individual test must be based upon whether it is being used in an epidemiological survey, as a screening device or in the detailed evaluation and confirmation of a final diagnosis.[2,6]

The present discussion assumes a fixed model; that is, patients must be classified as diseased or nondiseased. However, a different situation will exist and a different approach to the presentation and analysis is necessary if one considers a disease state as a dynamic condition. The real-life implication is in the individuals who are assigned to the nondiseased group but who

in fact may be "pre" or "burned out" cases of the disease. The development of sensitivity and specificity data for such a model requires long-term follow-up (5 to 10 years or longer) to establish or exclude the diagnosis. Several studies evaluating specific tests are being carried out in selected disease states, such as those in patients believed to be prediabetic, prehypertensive, or in those who have early obstructive airways disease.

Many of the nuclear medicine test procedures have fallen into disrepute because the concept responsible for the development and acceptance of the specialty—that of providing a safe, rapid, simple, screening test—has been forgotten and instead the specialty has been used as the definitive diagnostic test. Application of the concepts described in this paper to the evaluation of nuclear medicine test procedures should result in a clearer picture of the potential contribution of the specialty.

References

1. Armitage, P.: Statistical methods in medical research, New York, 1971, John Wiley & Sons, Inc.
2. Cochrane, A. L., and Holland, W. W.: Validation of screening procedures, Br. Med. Bull. **27:**3-8, 1971.
3. Fagen, T. J.: Nomogram for Bayes's theorem, N. Engl. J. Med. **293:**257, 1975.
4. Katz, M. A.: A probability graph describing the predictive value of a highly sensitive diagnostic test, N. Engl. J. Med. **291:**1115-1116, 1974.
5. Poulose, K. P., Reba, R. C., Cameron, J. C., et al.: The value and limitations of liver scanning for the detection of hepatic metastases in patients with cancer, J. Ind. Med. Assoc. **61:**199-205, 1973.
6. Sackett, D. L., and Holland, W. W.: Controversy in the detection of disease, Lancet **2:**357-359, 1975.
7. Vecchio, T. J.: Predictive value of a single diagnostic test in unselected populations, N. Engl. J. Med. **274:**1171-1173, 1966.

Suggested readings

1. Adelstein, S. J.: Determining the utility of nuclear medicine and radiologic procedures. In Goswitz, F. A., Andrews, G. A., Viamonte, M., Jr., editors: AEC Symposium 27, Clinical uses of radionuclides: critical comparison with other techniques, ORINS, Conf. 711101, Washington, D.C., 1972, U.S. Atomic Energy Commission Office, Information Services, pp. 662-667.
2. Feinstein, A. R.: Clinical judgment, Baltimore, 1967, The Williams & Wilkins Company.
3. Galen, R. S., and Gambino, S. R.: Beyond normality: the predictive value and efficiency of medical diagnosis, New York, 1975, John Wiley & Sons, Inc.
4. Lusted, L. B.: Introduction to medical decision making, Springfield, Ill., 1968, Charles C Thomas, Publisher.
5. Lusted, L. B.: Decision-making studies in patient management, N. Engl. J. Med. **284:**416-424, 1971.
6. McNeil, B. J., Keeler, E., and Adelstein, S. J.: Primer on certain elements of medical decision making, N. Engl. J. Med. **293:**211-215, 1975.
7. Murphy, E. A.: The logic of medicine, Baltimore, 1976, The Johns Hopkins University Press.
8. Schwartz, W. B., Gorry, G. A., Kassirer, J. P., and Essig, A.: Decision analysis and clinical judgment, Am. J. Med. **55:**459-472, 1973.
9. Wilson, J. M.: Current trends and problems in health screening, J. Clin. Pathol. **26:**555-563, 1973.