chapter

5 Assessing the Validity and Reliability of Diagnostic and Screening Tests

A normal individual is a person
Who has not been sufficiently examined.
– *Anonymous*

To understand how a disease is transmitted and develops and to provide appropriate and effective health care, it is necessary to distinguish between people in the population who have the disease and those who do not. This is an important challenge, both in the clinical arena, where patient care is the issue, and in the public health arena, where secondary prevention programs that involve early disease detection and intervention are being considered and where etiologic studies are being conducted to provide a basis for primary prevention. Thus, the quality of screening and diagnostic tests is a critical issue. Regardless of whether the test is a physical examination, a chest x-ray, an electrocardiogram, or a blood or urine assay, the same issue arises: How good is the test in separating populations of people with and without the disease in question? This chapter addresses the question of how we assess the quality of newly available screening and diagnostic tests to make reasonable decisions about their use and interpretation.

Biologic Variation of Human Populations

In using a test to distinguish between individuals with normal and abnormal results, it is important to understand how characteristics are distributed in human populations.

Figure 5–1 shows the distribution of tuberculin test results in a population. The size of the induration (area of hardness at the site of the injection) in millimeters is shown on the horizontal axis and the number of individuals is indicated on the vertical axis. A large group centers on the value of 0 mm—no induration—and another group centers near 20 mm of induration. This type of distribution, in which there are two peaks, is called a *bimodal curve*. The bimodal distribution permits the separation of individuals who had no prior experience with tuberculosis (people with no induration, seen on the left) from those who had prior experience with tuberculosis (those with about 20 mm of induration, seen on the right). Although some individuals fall into the "gray zone" in the center, and may belong to either curve, most of the population can be easily distinguished using the two curves. Thus, when a characteristic has a bimodal distribution, it is relatively easy to separate most of the population into two groups (e.g., ill and not ill, having a certain condition or abnormality and not having that condition or abnormality).

In general, however, most human characteristics are not distributed bimodally. Figure 5–2 shows the distribution of systolic blood pressures in a group of men. In this figure there is no bimodal curve; what we see is a *unimodal curve*—a single peak. Therefore, if we want to separate those in the group who are hypertensive from those who are not hypertensive, a cutoff level of blood pressure must be set above which people are designated hypertensive and below which they are designated normotensive. No obvious level of blood pressure distinguishes normotensive from hypertensive individuals. Although we could choose a cutoff for hypertension based on statistical considerations, we would ideally like to choose a cutoff on the basis of biologic information; that is, we would want to know that a pressure above the chosen cutoff level is associated with increased risk of subsequent disease, such as stroke, myocardial infarction, or

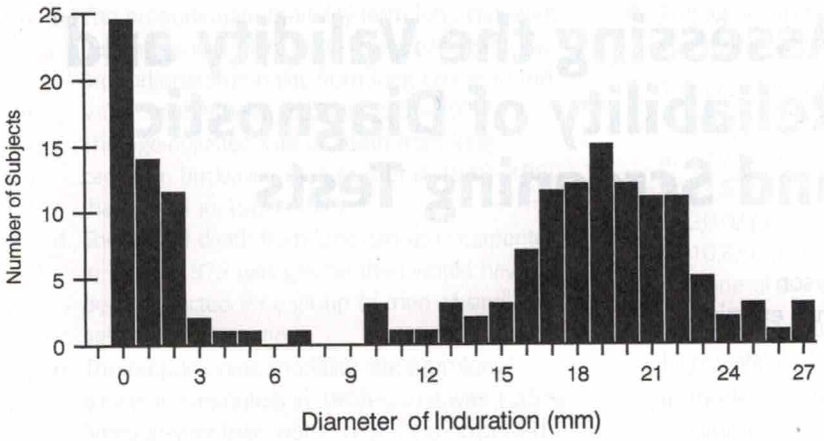


FIGURE 5–1 ▼ Distribution of tuberculin reactions. (Adapted from Edwards LB, Palmer CE, Magnus K: BCG Vaccination: Studies by the WHO Tuberculosis Research Office, Copenhagen. WHO Monograph No. 12. Geneva, WHO, 1953.)

subsequent mortality. Unfortunately, for many human characteristics, we do not have such information to serve as a guide in setting this level.

In either distribution—unimodal or bimodal—it is relatively easy to distinguish between the extreme values of abnormal and normal. With either type of curve, however, uncertainty remains about cases that fall into the gray zone.

Validity of Screening Tests

The *validity* of a test is defined as its ability to distinguish between who has a disease and who does not. Validity has two components: sensitivity and specificity. The *sensitivity* of the test is defined as the ability of the test to identify correctly those who *have* the disease. The *specificity* of the test is defined as the ability of the test to identify correctly those who *do not have* the disease.

Tests with Dichotomous Results (Positive or Negative)

Suppose we have a hypothetical population of 1,000 people, of whom 100 have a certain disease and 900 do not. A test is available that can yield either positive or negative results. We want to use this test to try to distinguish persons who have the disease from those who do not. The results obtained by applying the test to this population of 1,000 people are shown in Table 5–1.

How good was the test? First, how good was the test in correctly identifying those who had the disease? Table 5–1 indicates that of the 100 people with the disease, 80 were correctly identified as "positive" by the test, and a positive identification

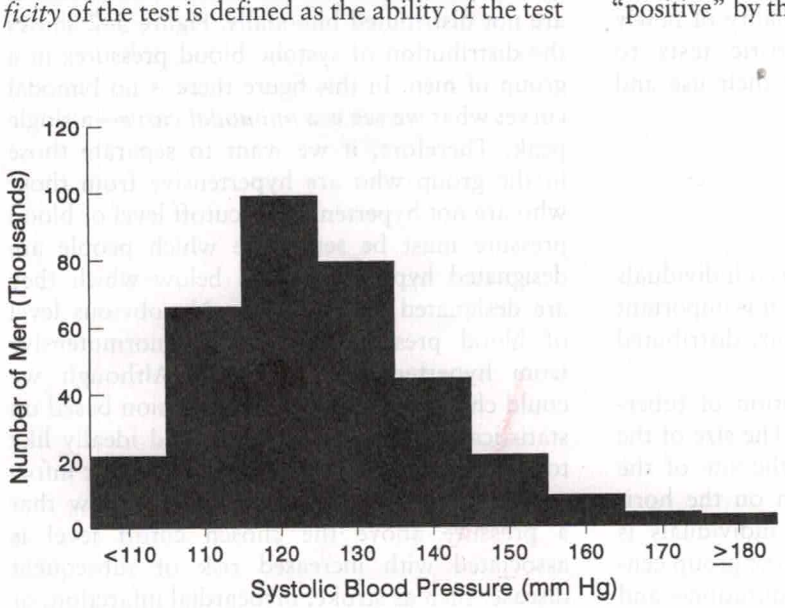


FIGURE 5–2 ▼ Distribution of systolic blood pressure for men screened for the Multiple Risk Factor Intervention Trial. (Data from Stamler J, Stamler R, Neaton JD: Blood pressure, systolic and diastolic, and cardiovascular risks: U.S. population data. Arch Intern Med 153:598–615, 1993.)

TABLE 5–1 Concept of the Sensitivity and Specificity of Screening Examinations

Example: Assume a population of 1,000 people, of whom 100 have a disease and 900 do not have the disease

Screening Test to Identify the 100 People with the Disease

Results of Screening	True Characteristics in the Population: *Disease*	*No Disease*	Total
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

$$\text{Sensitivity} = \frac{80}{100} = 80\%$$

$$\text{Specificity} = \frac{800}{900} = 89\%$$

was missed in 20. Thus, the *sensitivity* of the test, which is defined as the proportion of diseased people who were correctly identified as "positive" by the test, is 80/100 or 80%.

Second, how good was the test in correctly identifying those who did not have the disease? Looking again at Table 5–1, of the 900 people who did not have the disease, the test correctly identified 800 as "negative." The *specificity* of the test, which is defined as the proportion of nondiseased people who are correctly identified as negative by the test, is therefore 800/900 or 89%.

Note that to calculate the sensitivity and specificity of a test, we must know who "really" has the disease and who does not from another source than the test we are using. We are in fact comparing our test results with some "gold standard"—an external source of "truth" regarding the disease status of each individual in the population. Sometimes this truth may be the result of another test that has been in use, and sometimes it is the result of a more definitive, and often more invasive, test (e.g., cardiac catheterization or tissue biopsy). However, in real life, when we use a test to identify diseased and nondiseased persons in a population, we clearly do not know who has the disease and who does not. (If this were already established, testing would be pointless.) But to quantitatively assess the sensitivity and specificity of a test, we must have another source of truth with which to compare the test results.

Table 5–2 compares the results of a dichotomous test (results either positive or negative) with the actual disease status. Ideally, we would like all of the tested subjects to fall into the two cells shown in the upper left and lower right on the table: people with the disease who are correctly called "positive" by the test *(true positives)* and people without the disease who are correctly called "negative" by the test *(true negatives)*. Unfortunately, such is rarely if ever the case. Some people who do not have the disease are erroneously called "positive" by the test *(false positives)*, and some people with the disease are erroneously called "negative" *(false negatives)*.

Why are these issues important? When we conduct a screening program, we often have a large group of people who screened positive, including both people who really have the disease (true positives) and people who do not have the disease (false positives). The issue of *false positives* is important because all people who screened positive are brought back for more sophisticated and more expensive tests. Of the

TABLE 5–2 Comparison of the Results of a Dichotomous Test with Disease Status

Test Results	Population: *With Disease*	*Without Disease*
Positive	True positive (TP) = Have disease and have positive test	False positive (FP) = No disease, but have positive test
Negative	False negative (FN) = Have disease, but have negative test	True negative (TN) = No disease and have negative test

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Specificity} = \frac{TN}{TN + FP}$$

several problems that result, the first is a burden on the health care system. Another is the anxiety and worry induced in persons who have been told that they have tested positive. Considerable evidence indicates that many people who are labeled "positive" by a screening test never have that label completely erased, even if the results of a subsequent evaluation are negative. For example, children labeled "positive" in a screening program for heart disease were handled as handicapped by parents and school personnel even after being told that subsequent more definitive tests were negative. In addition, such individuals may be limited in regard to employment and insurability by erroneous interpretation of positive screening test results, even if subsequent tests fail to substantiate any positive finding.

Why is the problem of *false negatives* important? If a person has the disease but is erroneously informed that the test result is negative, and if the disease is a serious one for which effective intervention is available, the problem is indeed critical. For example, if the disease is a type of cancer that is curable only in its early stages, a false-negative result could represent a virtual death sentence. Thus, the importance of false-negative results depends on the nature and severity of the disease being screened for, the effectiveness of available intervention measures, and whether the effectiveness is greater if the intervention is administered early in the natural history of the disease.

Tests of Continuous Variables

So far we have discussed a test with only two possible results: positive or negative. But we often test for a continuous variable, such as blood pressure or blood glucose level, for which there is no "positive" or "negative" result. A decision must therefore be made in establishing a cutoff level above which a test result is considered positive and below which a result is considered negative. Let us consider the diagrams shown in Figure 5–3.

Figure 5–3*A* shows a population of 20 diabetics and 20 nondiabetics who are being screened using a blood sugar test whose scale is shown along the vertical axis from high to low. The diabetics are represented by solid circles and the nondiabetics by cross-hatched circles. We see that although blood sugar levels tend to be higher in diabetics than in nondiabetics, no level clearly separates the two groups; there is some overlap of diabetics and nondiabetics at every blood sugar level. Nevertheless, we must select a cutoff point so that those whose results fall above the cutoff can be called "positive," and can be called back for further testing, and those whose results fall below that point are called "negative," and are not called back for further testing.

Suppose a relatively high cutoff level is chosen (Fig. 5–3*B*). Clearly, many of the diabetics will not be identified as positive; on the other hand, most of the nondiabetics will be correctly identified as negative. If these results are distributed on a 2×2 table, the sensitivity of the test using this cutoff level will be 25% (5/20) and the specificity will be 90% (18/20).

What if a low cutoff level is chosen (Fig. 5–3*C*)? Very few diabetics would be misdiagnosed. What then is the problem? A large proportion of the nondiabetics are now identified as positive by the test. As seen in the 2×2 table, the sensitivity is now 85% (17/20), but the specificity is only 30% (6/20).

The difficulty is that in the real world, no vertical line separates the diabetics and nondiabetics, and they are, in fact, mixed together (Fig. 5–3*D*); in fact, they are not even distinguishable by solid or cross-hatched circles (Fig. 5–3*E*). So if a high cutoff level is used (Fig. 5–3*F*), all those with results below the line will be assured they do not have the disease and will not be followed further; if the low cutoff is used (Fig. 5–3*G*), all those with results above the line will be brought back for further testing.

Figure 5–4*A* shows actual data regarding the distribution of blood sugar levels in diabetics and nondiabetics. Suppose we were to screen this population. If we decide to set the cutoff level so that we identify all of the diabetics (100% sensitivity), we could set the level at 80 mg/dl (Figure 5–4*B*). The problem is, however, that in so doing we will also call many of the nondiabetics positive (very low specificity). On the other hand, if we set the level at 200 mg/dl (Figure 5–4*C*) so that we call all the nondiabetics negative (100% specificity), we now miss many of the true diabetics (very low sensitivity). Thus, there is a trade-off between sensitivity and specificity: if we increase the sensitivity by lowering the cutoff level, we decrease the specificity; if we increase the specificity by raising the cutoff level, we decrease the sensitivity. To quote an unknown sage: "There is no such thing as a free lunch."

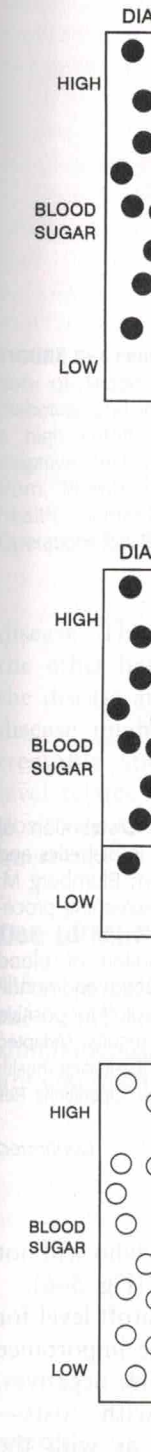


FIGURE 5–3 ▼
different cuto

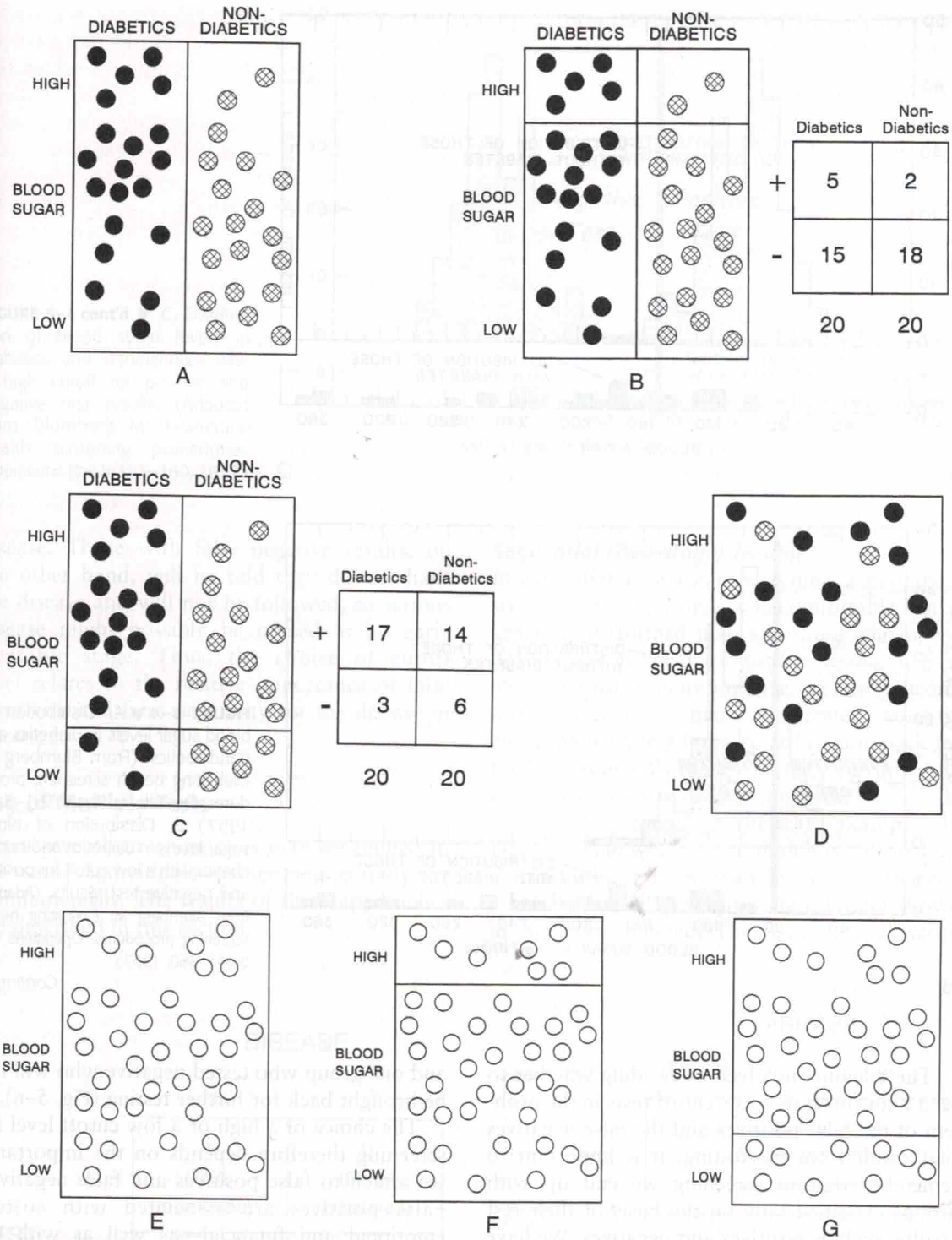
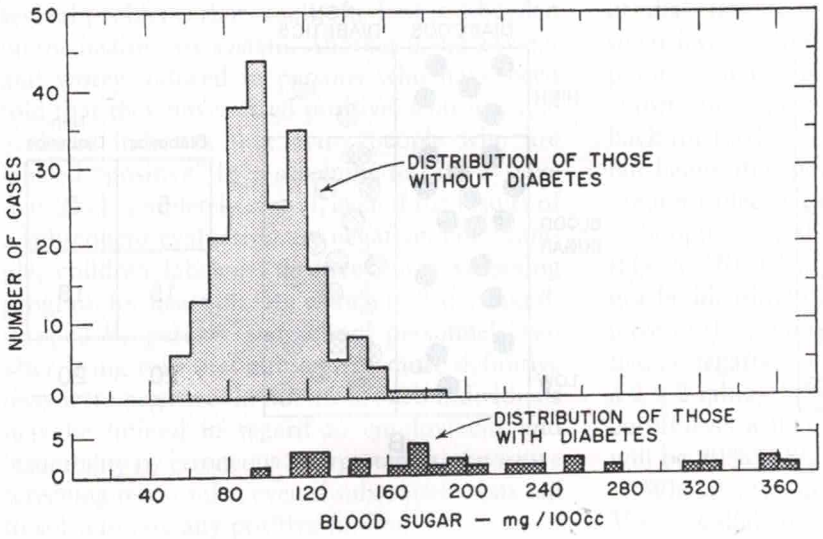
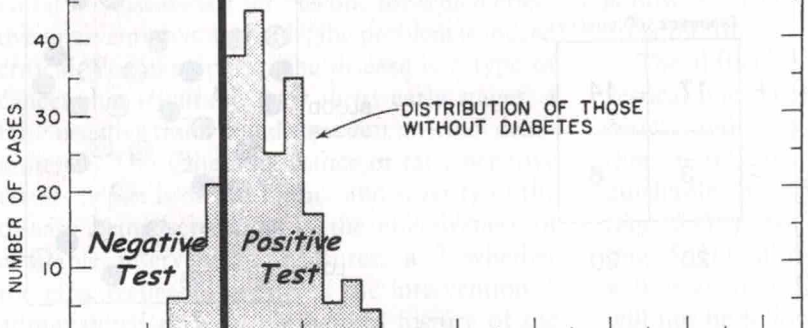


FIGURE 5–3 ▼ *A–G*, Screening for diabetes in a hypothetical population with a prevalence of 50%. Effects of choosing different cutoff levels for a positive test. (See text.)



A



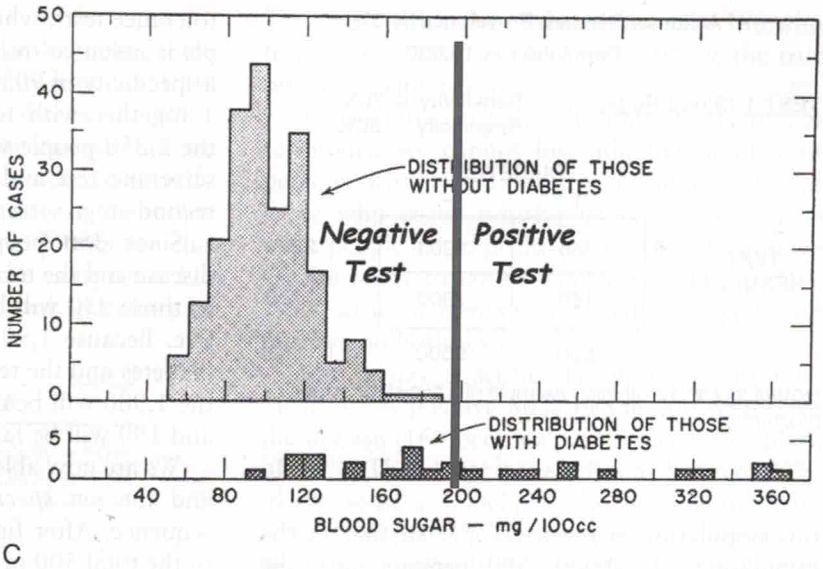
B

FIGURE 5–4 ▼ *A*, Distribution of blood sugar levels in diabetics and nondiabetics. (From Blumberg M: Evaluating health screening procedures. Operations Res 5:351–360, 1957.) *B*, Distribution of blood sugar levels in diabetics and nondiabetics with a low cutoff for positive and negative test results. (Adapted from Blumberg M: Evaluating health screening procedures. Operations Res 5:351–360, 1957.)

Continued

The dilemma involved in deciding whether to set a high cutoff or a low cutoff rests in the problem of the false positives and the false negatives that result from the testing. It is important to remember that in screening we end up with groups classified only on the basis of their test results, such as positives and negatives. We have no information regarding their true disease status, which, of course, is the reason for the screening. In effect, the results yield not four groups, as seen in Figure 5–5, but rather two groups: one group of people who tested positive and who will be brought back for additional examinations and one group who tested negative who will not be brought back for further testing (Fig. 5–6).

The choice of a high or a low cutoff level for screening therefore depends on the importance we attach to false positives and false negatives. False positives are associated with costs—emotional and financial—as well as with the difficulty of "delabeling" a person who tests positive and is later found not to have the disease. In addition, false positive results pose a major burden to the health care system in that a large group of people need to be brought back for a retest, when only a few of them may have the



C

FIGURE 5–4 cont'd ▼ *C*, Distribution of blood sugar levels in diabetics and nondiabetics with a high cutoff for positive and negative test results. (Adapted from Blumberg M: Evaluating health screening procedures. Operations Res 5:351–360, 1957.)

disease. Those with false negative results, on the other hand, will be told they do not have the disease and will not be followed, so serious disease might possibly be missed at an early treatable stage. Thus, the choice of cutoff level relates to the relative importance of false positivity and false negativity for the disease in question.

Use of Multiple Tests

Often several screening tests may be applied in the same individuals—either sequentially or simultaneously. The results of these approaches are described in this section.

Sequential (Two-stage) Testing

In sequential or two-stage screening, a less expensive, less invasive, or less uncomfortable test is generally performed first, and those who screen positive are recalled for further testing with a more expensive, more invasive, or more uncomfortable test, which may have greater sensitivity and specificity. It is hoped that bringing back for further testing only those who screen positive will reduce the problem of false positives.

Consider the hypothetical example in Figure 5–7, in which a population is screened for diabetes using a test with a sensitivity of 70% and a specificity of 80%. How are the data shown in this table obtained? The disease prevalence in

TEST	DISEASE +	DISEASE −
+	a (True positives)	b (False positives)
−	c (False negatives)	d (True negatives)

FIGURE 5–5 ▼ Diagram showing four possible groups resulting from screening with a dichotomous test.

TEST	DISEASE +	DISEASE −
+	a + b (All people with positive tests)	
−	c + d (All people with negative tests)	

FIGURE 5–6 ▼ Diagram grouping all people with positive test results and all people with negative test results on screening.

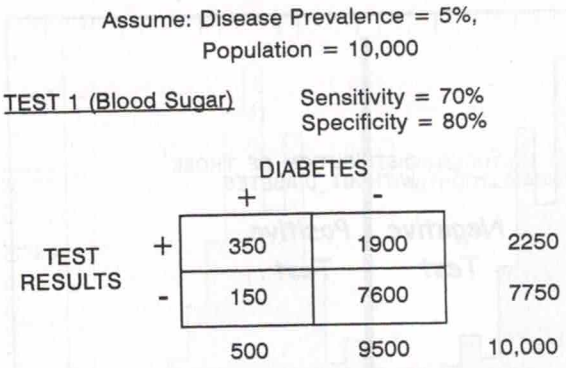


FIGURE 5–7 ▼ Hypothetical example of a two-stage screening program: I.

this population is given as 5%, so that in the population of 10,000, 500 persons have the disease. With a sensitivity of 70%, the test will correctly identify 350 of the 500 people who have the disease. With a specificity of 80%, the test will correctly identify as nondiabetic 7,600 of the 9,500 people who are free of diabetes; however, 1,900 of these 9,500 will have positive results. Thus a total of 2,250 people will test positive and will be brought back for a second test. (Remember that in real life we do not have the vertical line separating diabetics and nondiabetics, and we do not know that 350 of the 2,250 have diabetes.)

Now those 2,250 people are brought back and screened using a second test (such as a glucose tolerance test), which for purposes of this example is assumed to have a sensitivity of 90% and a specificity of 90%. Figure 5–8 again shows test 1 together with test 2, which deals only with the 2,250 people who tested positive in the first screening test and have been brought back for second-stage screening.

Since 350 people (of the 2,250) have the disease and the test has a sensitivity of 90%, 315 of those 350 will be correctly identified as positive. Because 1,900 (of the 2,250) do not have diabetes and the test specificity is 90%, 1,710 of the 1,900 will be correctly identified as negative and 190 will be false positives.

We are now able to calculate the *net sensitivity* and the *net specificity* of using both tests in sequence. After finishing both tests, 315 people of the total 500 in this population of 10,000 will have been correctly called positive: 315/500 = 63% *net sensitivity.* Thus, there is a loss in net sensitivity by using both tests. To calculate *net specificity,* note that 7,600 people of the 9,500 in this population who do not have diabetes were correctly called negative in the first-stage screening and were not tested further; an additional 1,710 of those 9,500 nondiabetics were correctly called negative in the second-stage screening. Thus a total of 7,600 + 1,710 of the 9,500 nondiabetics were correctly called negative: 9,310/9,500 = 98% *net specificity.* Thus, use of both tests has resulted in a gain in *net specificity.*

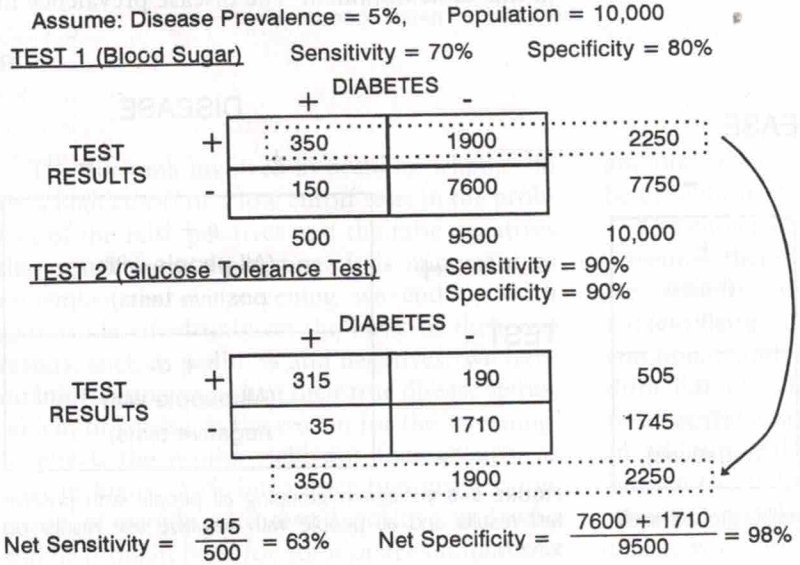


FIGURE 5–8 ▼ Hypothetical example of a two-stage screening program: II.

Simultaneous Testing

Let us now turn to the use of simultaneous tests. Let us assume that, in a population of 1000 people, the prevalence of a disease is 20%. Therefore, 200 people have the disease, but we do not know who they are. In order to identify the 200 people who have this disease, we screen this population of 1,000 using 2 tests for this disease, test A and test B, at the same time. Let us assume that the sensitivity and specificity of the two tests are as follows:

Test A	Test B
Sensitivity = 80% Specificity = 60%	Sensitivity = 90% Specificity = 90%

Net Sensitivity Using Two Simultaneous Tests

The first question we ask is, "What is the *net sensitivity* using test A and test B *simultaneously*?" To be considered positive and therefore included in the numerator for net sensitivity for two tests used simultaneously, a person must be identified as positive by test A, test B, or both tests.

To calculate net sensitivity, let us first consider the results of screening with test A sensitivity is 80%: Of the 200 people who have the disease, 160 test positive (Table 5–3). In Figure 5–9*A*, the oval represents the 200 people who have the disease. The circle within the oval represents the 160 who test positive with test A. These 160 are the true positives using test A.

Consider next the results of screening with test B whose sensitivity is 90% (Table 5–4). Of the 200 people who have the disease, 180 test positive by test B. In Figure 5–9*B*, the oval again represents the 200 people who have the disease. The circle within the oval represents the 180 who test positive with test B. These 180 are the true positives using test B.

In order to calculate the numerator for net sensitivity, we cannot just add the number of persons who tested positive using test A and those who tested positive using test B because some people tested positive on both tests, and we do not want to count them twice (Fig. 5–9*C*). How do we determine how many people tested positive on both tests?

Test A has a sensitivity of 80% and thus identifies as positive 80% of the 200 who have the disease (160 people). Test B has a sensitivity of 90%. Therefore, it identifies as positive 90% of the same 160 people who are identified by test A (144 people). Thus, when tests A and B are used simultaneously, 144 people are identified as positive by both tests (Fig. 5–9*D*).

Recall that test A correctly identified 160 people with the disease as positive. Because 144 of them were identified by both tests, 160 – 144, or 16 people, were correctly identified *only* by test A.

Test B correctly identified 180 of the 200 people with the disease as positive. Because 144 of them were identified by both tests, 180 – 144, or 36 people, were correctly identified *only* by test B. Thus, as seen in Figure 5–9*E*, the net sensitivity using tests A and B simultaneously

$$= \frac{16 + 144 + 36}{200} = \frac{196}{200} = 98\%$$

Net Specificity Using Two Simultaneous Tests

The next question is, "What is the *net specificity* using test A and test B simultaneously?" To be included in the numerator for net specificity for two tests used simultaneously, a person must be identified as negative by both tests. In order to calculate the numerator for net specificity, we

TABLE 5–3 Results of Screening With Test A

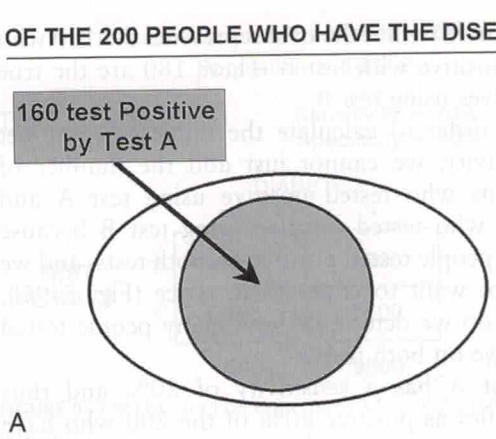
Results of Screening	Population	
	Disease	*No Disease*
Positive	(160)	320
Negative	40	480
Total	200	800

Sensitivity = 80%
Specificity = 60%

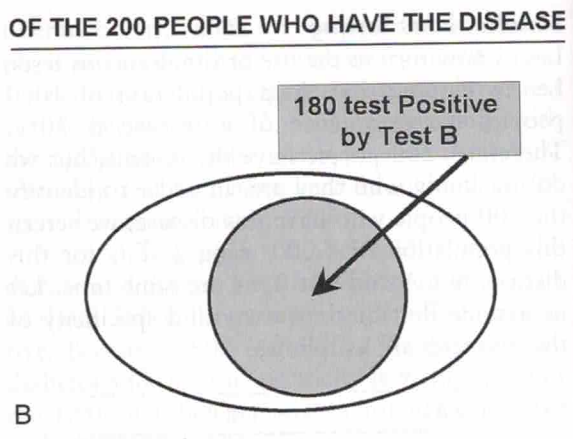
TABLE 5–4 Results of Screening With Test B

Results of Screening	Population	
	Disease	*No Disease*
Positive	(180)	80
Negative	20	720
Total	200	800

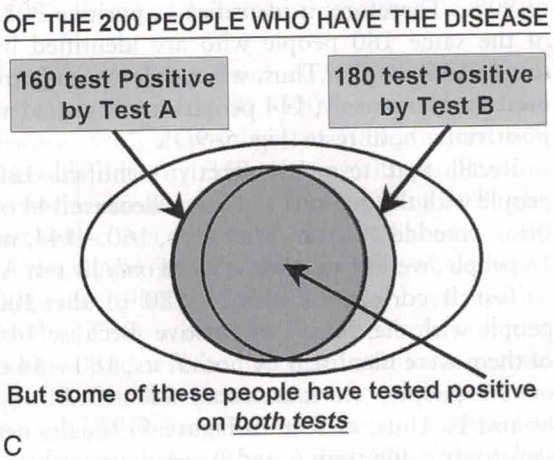
Sensitivity = 90%
Specificity = 90%



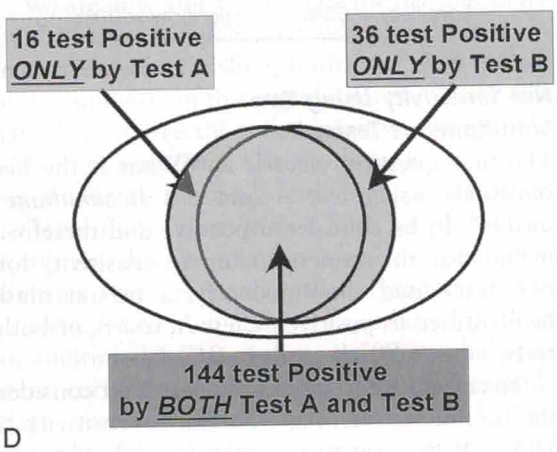
A



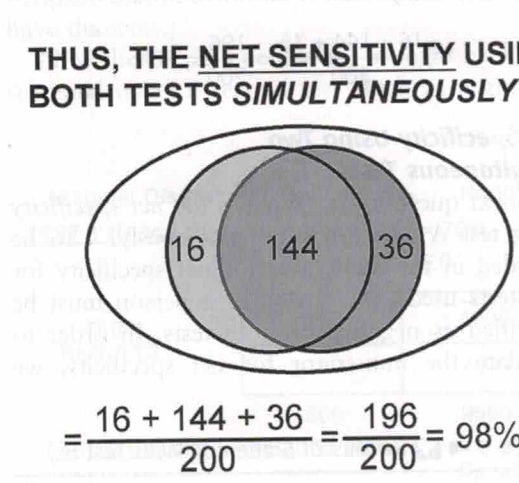
B



C



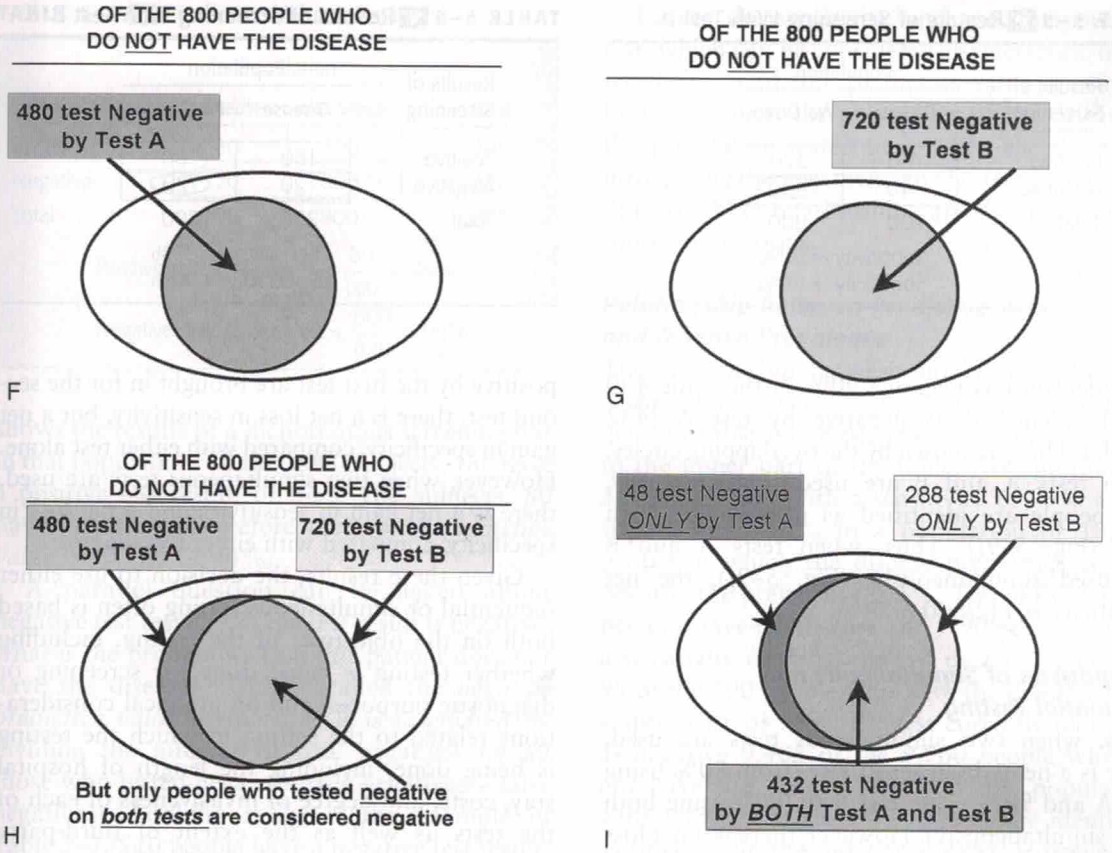
D



E

FIGURE 5–9 ▼ Hypothetical example of simultaneous testing: *A–E:* Net sensitivity (see text).

Continued



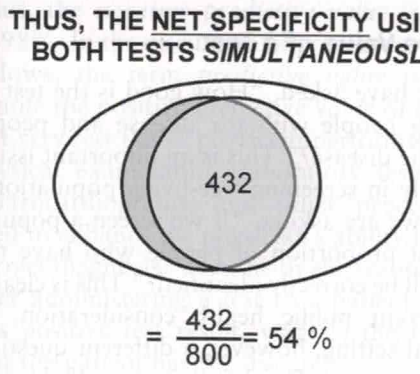


FIGURE 5–9 cont'd ▼ F–J: Net specificity (see text).

therefore need to determine how many people had negative results on both tests. How do we do this?

Test A has a specificity of 60% and thus correctly identifies 60% of the 800 who do not have the disease (480 people) (Table 5–5). In Figure 5–9*F*, the oval represents the 800 people who do not have the disease. The circle within the oval represents the 480 people who test negative on test A. These are the true negatives using test A.

Test B has a specificity of 90% and thus identifies as negative 90% of the 800 people who do not have the disease (720 people) (Table 5–6 and Figure 5–9*G*). However, to be called negative in simultaneous tests, only people who test negative on both tests are considered to have had negative results (Fig. 5–9*H*). Test B

TABLE 5–5 Results of Screening With Test A

Results of Screening	Population	
	Disease	*No Disease*
Positive	160	320
Negative	40	480
Total	200	800

Sensitivity = 80%
Specificity = 60%

TABLE 5–6 Results of Screening With Test B

Results of Screening	Population	
	Disease	*No Disease*
Positive	180	80
Negative	20	720
Total	200	800

Sensitivity = 90%
Specificity = 90%

also identifies as negative 90% of the same 480 people identified as negative by test A (432 people). Thus, as shown by the overlapping circles, when tests A and B are used simultaneously, 432 people are identified as negative by both tests (Fig. 5–9*I*). Thus, when tests A and B are used simultaneously, (Fig. 5–9*J*), the net specificity = 432/800 = 54%.

Comparison of Simultaneous and Sequential Testing

Thus, when two simultaneous tests are used, there is a net gain in sensitivity (from 80% using test A and 90% using test B to 98% using both tests simultaneously). However, there is a net loss in specificity (net specificity = 54%) compared to using either test alone (specificity of 60% using test A and 90% using test B).

In a clinical setting, multiple tests are often used simultaneously. For example, a patient admitted to a hospital may have an array of tests performed at the time of admission. When multiple tests are used simultaneously to detect a specific disease, the individual is generally considered to have tested "positive" if he or she has a positive result on any one or more of the tests. The individual is considered to have tested "negative" if he or she tests negative on all of the tests. The effects of such a testing approach on sensitivity and specificity differ from those that result from sequential testing. In sequential testing when we retest those who tested positive on the first test, there is a loss in net sensitivity and a gain in net specificity. In simultaneous testing, because an individual who tests positive on *any* one or multiple tests is considered positive, there is a gain in net sensitivity. However, to be considered negative, a person would have to test negative on *all* the tests performed. As a result, there is a loss in net specificity.

In summary, as we have seen previously, when two sequential tests are used and those who test positive by the first test are brought in for the second test, there is a net loss in sensitivity, but a net gain in specificity, compared with either test alone. However, when two simultaneous tests are used, there is a net gain in sensitivity and a net loss in specificity, compared with either test alone.

Given these results, the decision to use either sequential or simultaneous testing often is based both on the objectives of the testing, including whether testing is being done for screening or diagnostic purposes, and on practical considerations related to the setting in which the testing is being done, including the length of hospital stay, costs, and degree of invasiveness of each of the tests as well as the extent of third-party insurance coverage.

Predictive Value of a Test

So far, we have asked, "How good is the test at identifying people with the disease and people without the disease?" This is an important issue, particularly in screening free-living populations. In effect, we are asking, "If we screen a population, what proportion of people who have the disease will be correctly identified?" This is clearly an important public health consideration. In the clinical setting, however, a different question may be important for the physician: If the test results are positive in this patient, what is the probability that this patient has the disease? This is called the *positive predictive value* of the test. In other words, what proportion of patients who test positive actually have the disease in question? To calculate the predictive value, we divide the number of true positives by the total number who tested positive (true positives + false positives).

Let us return to the example shown in Table 5–1, in which a population of 1,000 persons is screened. As seen in Table 5–7, a 2 × 2 table

TABLE 5–7 Predictive Value of a Test

	Population		
Test Results	*Disease*	*No Disease*	Total
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

$$\text{Positive predictive value} = \frac{80}{180} = 44\%$$

$$\text{Negative predictive value} = \frac{800}{820} = 98\%$$

shows the results of a dichotomous screening test in that population. Of the 1,000 subjects, 180 have a positive test result; of these 180 subjects, 80 have the disease. Therefore, the *positive predictive value* is 80/180 or 44%.

A parallel question can be asked about negative test results: "If the test result is negative, what is the probability that this patient does not have the disease?" This is called the *negative predictive value* of the test. It is calculated by dividing the number of true negatives by all those who tested negative (true negatives + false negatives). Looking again at the example in Table 5–7, 820 people have a negative test result, and of these, 800 do not have the disease. Thus, the *negative predictive value* is 800/820 or 98%. In the discussion of predictive value that follows, the term *predictive value* is used to denote the positive predictive value of the test.

Every test that a physician performs—history, physical examination, laboratory tests, x-rays, electrocardiograms, and other procedures—is used to enhance the physician's ability to make a correct diagnosis. What he or she wants to know after administering a test to a patient is: Given this positive test result, what is the likelihood that the patient has the disease?"

Unlike the sensitivity and specificity of the test, which can be considered characteristic of the test being used, the predictive value is affected by two factors: the prevalence of the disease in the population tested and, when the disease is infrequent, the specificity of the test being used. Both of these relationships are discussed in the following sections.

Relationship between Predictive Value and Disease Prevalence

The relationship between predictive value and *disease prevalence* can be seen in the example given in Table 5–8. First, let us direct our attention to the upper part of the table. Assume that we are using a test with a sensitivity of 99% and a specificity of 95% in a population of 10,000 people in which the disease prevalence is 1%. Because the prevalence is 1%, 100 of the 10,000 persons have the disease and 9,900 do not. With a sensitivity of 99%, the test correctly identifies 99 of the 100 people who have the disease. With a specificity of 95%, the test correctly identifies as negative 9,405 of the 9,900 people who do not have the disease. Thus, in this population with a 1% prevalence, 594 people are identified as positive by the test (99 + 495). However, of these 594 people, 495 (83%) are false positives and the positive predictive value is therefore 99/594 or only 17%.

Let us now apply the same test—with the same sensitivity and specificity—to a population with a higher disease prevalence, 5%, as seen in the lower part of Table 5–8. Using calculations similar to those used in the upper part of the table, the positive predictive value is now 51%. Thus, the higher prevalence in the screened population has led to a marked increase in the positive predictive value using the same test. Figure 5–10 shows the relationship between

TABLE 5–8 Relationship of Disease Prevalence to Positive Predictive Value

	Example: Sensitivity = 99%, specificity = 95%				
Disease Prevalence	Test Results	Sick	Not Sick	Totals	Positive Predictive Value
1%	+	99	495	594	$\frac{99}{594} = 17\%$
	−	1	9,405	9,406	
	Totals	100	9,900	10,000	
5%	+	495	475	970	$\frac{495}{970} = 51\%$
	−	5	9,025	9,030	
	Totals	500	9,500	10,000	

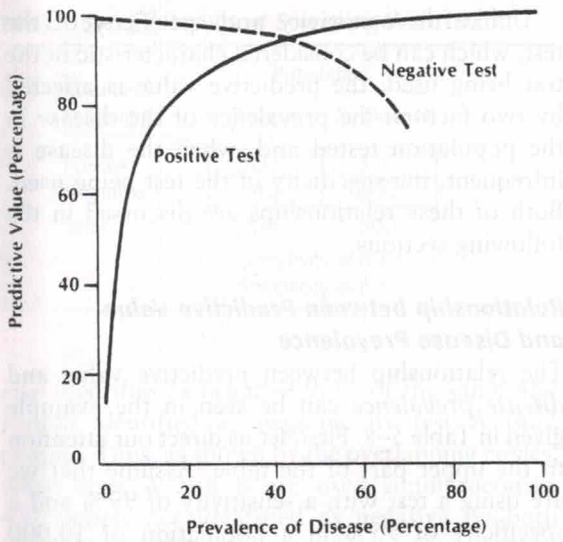


FIGURE 5–10 ▼ Relationship between disease prevalence and predictive value in a test with 95% sensitivity and 95% specificity. (From Mausner JS, Kramer S: Mausner and Bahn Epidemiology: An Introductory Text. Philadelphia, WB Saunders, 1985, p 221.)

disease prevalence and predictive value. Clearly, most of the gain in predictive value occurs with increases in prevalence at the lowest rates of disease prevalence.

Why should we be concerned about the relationship between predictive value and disease prevalence? As we have seen, the higher the prevalence, the higher the predictive value. Therefore, a screening program is most productive and efficient if it is directed to a high-risk target population. Screening a total population for a relatively infrequent disease can be very wasteful of resources and may yield few previously undetected cases relative to the amount of effort involved. However, if a high-risk subset can be identified and screening can be directed to this group, the program is likely to be far more productive. In addition, a high-risk population may be more motivated to participate in such a screening program and more likely to take recommended action if their screening results are positive.

The relationship between predictive value and disease prevalence also shows that the results of any test must be interpreted in the context of the prevalence of the disease in the population from which the subject originates. An interesting example is seen with the measurement of the α-fetoprotein (AFP) level in amniotic fluid for prenatal diagnosis of spina bifida. Figure 5–11 shows the distribution of AFP levels in amniotic fluid in normal pregnancies and in pregnancies in which the fetus has spina bifida, which is a neural tube defect. Although the distribution is bimodal, there is a range in which the curves overlap, and within that range, it may not always be clear to which curve the mother and fetus belong. Sheffield and coworkers[1] reviewed the literature and constructed artificial populations of 10,000 women screened for amniotic fluid AFP to identify fetuses with spina bifida. They created two populations: one at high risk for spina bifida and the other at normal risk.

Table 5–9 shows the calculations for both high-risk and low-risk women. Which women are at high risk for having a child with spina bifida? It is known that women who have previously had a child with a neural tube defect are at increased risk because the defect is known to repeat in siblings. In these calculations, the positive predictive value is found to be 82.9%. Which women are at low risk, but would still have an amniocentesis? These are older women who are

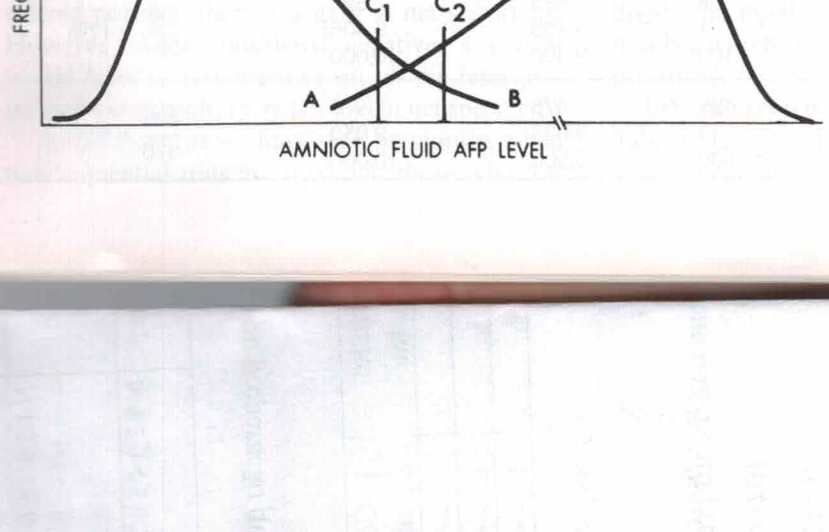


FIGURE 5–11 ▼ Amniotic fluid α-fetoprotein (AFP) levels in normal subjects and subjects with spina bifida. (From Sheffield LJ, Sackett DL, Goldsmith CH, et al: A clinical approach to the use of predictive values in the prenatal diagnosis of neural tube defects. Am J Obstet Gynecol 145:319–324, 1983.)

C

TABLE 5–9 ▼ Calcul
in High- and Low-Risk

High-risk women
Low-risk women

*Spina bifida or encephaloc
From Sheffield LJ, Sackett D
defects. Am J Obstet Gynec

undergoing amnioc
about possible Dow
defect associated wi
maternal age. The r
is not related to m
are not at increased
spina bifida. The c
the same test for A
risk women, the po
test is only 41.7%,
in a high-risk group

Thus, we see that
different predictive
to a high-risk (high
low-risk (low prev
clear clinical implica
decision to terminat
may formulate adv
basis of the test re
result may be interp
whether the woma
risk or low-risk wo
in the positive p
Consequently, by it
sufficient to serve a
account the other c

The following
importance of this

The head of a fire
university cardiolo
physician had rea
medical journal re
electrocardiograph
of serious, genera
heart disease. On
department physi

TABLE 5–9 Calculations of Predictive Values for Neural Tube Defects (NTD)* for α-Fetoprotein (AFP) Test in High- and Low-Risk Women

		Pregnancy Outcome			
	AFP Test	*NTD*	*Normal*	Totals	Predictive Value (%)
High-risk women	Abnormal	87	18	105	82.9
	Normal	13	9,882	9,895	99.9
	Totals	100	9,900	10,000	
Low-risk women	Abnormal	128	179	307	41.7
	Normal	19	99,674	99,693	99.98
	Totals	147	99,853	100,000	

*Spina bifida or encephalocele.

From Sheffield LJ, Sackett DL, Goldsmith CH, et al: A clinical approach to the use of predictive values in the prenatal diagnosis of neural tube defects. Am J Obstet Gynecol 145:319–324, 1983.

undergoing amniocentesis because of concern about possible Down syndrome or some other defect associated with pregnancy at an advanced maternal age. The risk of spina bifida, however, is not related to maternal age, so these women are not at increased risk for having a child with spina bifida. The calculation shows that, using the same test for AFP as was used for the high-risk women, the positive predictive value of the test is only 41.7%, considerably less than it was in a high-risk group.

Thus, we see that the same test can have a very different predictive value when it is administered to a high-risk (high prevalence) population or to a low-risk (low prevalence) population. This has clear clinical implications: A woman may make a decision to terminate a pregnancy and a physician may formulate advice to such a woman on the basis of the test results. However, the same test result may be interpreted differently, depending on whether the woman comes from a pool of high-risk or low-risk women, which will be reflected in the positive predictive value of the test. Consequently, by itself, the test result may not be sufficient to serve as a guide without taking into account the other considerations just described.

The following true examples highlight the importance of this issue:

> The head of a firefighters' union consulted a university cardiologist because the fire department physician had read an article in a leading medical journal reporting that a certain electrocardiographic finding was highly predictive of serious, generally unrecognized, coronary heart disease. On the basis of this article, the fire department physician was disqualifying many young, able-bodied firefighters from active duty. The cardiologist read the paper and found that the study had been carried out in hospitalized patients.

What was the problem? Because hospitalized patients have a much higher prevalence of heart disease than does a group of young firefighters, the fire department physician had erroneously taken the high predictive value obtained in studying a high-prevalence population and inappropriately applied it to a low-prevalence population of healthy firefighters, in whom the same test would actually have a much lower predictive value.

Another example:

> A physician visited his general internist for a regular annual medical examination, which included a stool examination for occult blood. One of the three stool specimens examined in the test was positive. The internist told his physician-patient that the result was of no significance because he regularly encountered many false-positive test results in his busy practice. The test was repeated, and all three stool specimens were now negative. Nevertheless, sensing his patient's lingering concerns, the internist referred his physician-patient to a gastroenterologist. The gastroenterologist said, "In my experience, the positive stool finding is serious. Such a finding is almost always associated with pathologic gastrointestinal disorders. The subsequent negative test results mean nothing, because you could have a tumor that only bleeds intermittently."

Who was correct in this episode? The answer is that both the general internist and the

gastroenterologist were correct. The internist gave his assessment of predictive value based on his experience in his general medical practice—a population with a low prevalence of serious gastrointestinal disease. On the other hand, the gastroenterologist gave his assessment of the predictive value of the test based on his experience in his referral practice—a practice in which most patients are referred because of a likelihood of serious gastrointestinal illness—a high-prevalence population.

Relationship between Predictive Value and Specificity of the Test

A second factor that affects the predictive value of a test is the *specificity* of the test. Examples of this are shown first in graphical form and then in tabular form. Figure 5–12*A* through *D* diagrams the results of screening a population; however, the 2×2 tables in these figures differ from those in earlier figures: The size of each cell is proportional to the population it represents. In each figure the cells that represent persons who tested positive are

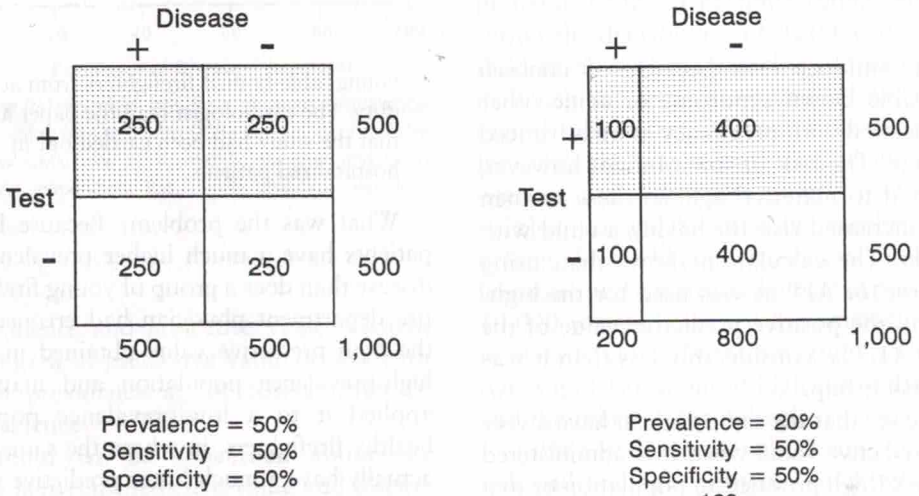


FIGURE 5–12 ▼ Relationship of specificity to predictive value (PV).

TABLE 5–10 ▼

Specificity
70%
95%

shaded gray; th
calculating the

Figure 5–12
population tha
population of
prevalence is
disease and 50
we also assum
used has a sens
50%. Because
250 of these
value is 250/50

Fortunately,
is much lower t
with relatively
Figure 5–12*B* a
(although even
prevalence for
and the specif
200 of the 1,00
vertical line sep
persons is shift
is now calculat

Given that
with the lower
the predictive
on predictive v
ity of the test?
when we leave
specificity at 5
90%. The pre
31%, a modes

What if, inst
the test, we inc
shows the resu
and sensitivity
increased to 9
100/180 or 56
resulted in a m
value than did

TABLE 5–10 Relationship of Specificity to Predictive Value

		Example: Prevalence = 10%, sensitivity = 100%			
Specificity	Test Results	Sick	Not Sick	Totals	Predictive Value
70%	+	1,000	2,700	3,700	$\frac{1{,}000}{3{,}700} = 27\%$
	–	0	6,300	6,300	
	Totals	1,000	9,000	10,000	
95%	+	1,000	450	1,450	$\frac{1{,}000}{1{,}450} = 69\%$
	–	0	8,550	8,550	
	Totals	1,000	9,000	10,000	

shaded gray; these are the cells that will be used in calculating the positive predictive value.

Figure 5–12*A* presents the baseline screened population that is used in our discussion: a population of 1,000 people in whom the prevalence is 50%; thus, 500 people have the disease and 500 do not. In analyzing this figure, we also assume that the screening test that was used has a sensitivity of 50% and a specificity of 50%. Because 500 people tested positive, and 250 of these have the disease, the predictive value is 250/500 or 50%.

Fortunately, the prevalence of most diseases is much lower than 50%; we are generally dealing with relatively infrequent diseases. Therefore, Figure 5–12*B* assumes a lower prevalence of 20% (although even this would be an unusually high prevalence for most diseases). Both the sensitivity and the specificity remain at 50%. Now only 200 of the 1,000 people have the disease, and the vertical line separating diseased from nondiseased persons is shifted to the left. The predictive value is now calculated as 100/500 or 20%.

Given that we are screening a population with the lower prevalence rate, can we improve the predictive value? What would be the effect on predictive value if we increased the sensitivity of the test? Figure 5–12*C* shows the results when we leave the prevalence at 20% and the specificity at 50% but increase the sensitivity to 90%. The predictive value is now 180/580 or 31%, a modest increase.

What if, instead of increasing the sensitivity of the test, we increase its specificity? Figure 5–12*D* shows the results when prevalence remains 20% and sensitivity remains 50%, but specificity is increased to 90%. The predictive value is now 100/180 or 56%. Thus, an increase in specificity resulted in a much greater increase in predictive value than did the same increase in sensitivity.

Why does specificity have a greater effect than sensitivity on predictive value? The answer becomes clear by examining these figures. Because we are dealing with infrequent diseases, most of the population falls to the right of the vertical line. Consequently, any change to the right of the vertical line affects a greater number of people than would a comparable change to the left of the line. Thus, a change in specificity has a greater effect on predictive value than does a change in sensitivity. If we were dealing with a high-prevalence disease, the situation would be different.

The effect of changes in specificity on predictive value is also seen in Table 5–10 in a form similar to that used in Table 5–8. As seen in this example, even with 100% sensitivity, a change in specificity from 70% to 95% has a dramatic effect on the positive predictive value.

Reliability (Repeatability) of Tests

Let us consider another aspect of assessing diagnostic and screening tests—the question of whether a test is reliable or repeatable. Can the results obtained be replicated if the test is repeated? Clearly, regardless of the sensitivity and specificity of a test, if the test results cannot be reproduced, the value and usefulness of the test are minimal. The rest of this chapter focuses on the reliability or repeatability of diagnostic and screening tests. The factors that contribute to the variation between test results are discussed first: intrasubject variation (variation within individual subjects) and interobserver variation (variation between those reading the test results).

Intrasubject Variation

The values obtained in measuring many human characteristics often vary over time, even during a short period. Table 5–11 shows changes in

TABLE 5–11 Examples Showing Variation in Blood Pressure Readings During a 24-Hour Period

Blood Pressure (mm Hg)	Female Aged 27 yr	Female Aged 62 yr	Male Aged 33 yr
Basal	110/70	132/82	152/109
Lowest hour	86/47	102/61	123/78
Highest hour	126/79	172/94	153/107
Casual	108/64	155/93	157/109

From Richardson DW, Honour AJ, Fenton GW, et al: Variation in arterial pressure throughout the day and night. Clin Sci 26:445, 1964.

blood pressure readings over a 24-hour period in three individuals. Variability over time is considerable. This, as well as the conditions under which certain tests are conducted (e.g., postprandially or postexercise, at home or in a physician's office), clearly can lead to different results in the same individual. Therefore, in evaluating any test result, it is important to consider the conditions under which the test was performed, including the time of day.

Intraobserver Variation

Sometimes variation occurs between two observations made by the same observer. For example, a radiologist who reads the same group of x-rays at two different times may read one or more of the x-rays differently the second time. Tests and examinations differ in the degree to which subjective factors enter into the observer's conclusions, and the greater the subjective element in the reading, the greater the intraobserver variation in readings is likely to be (Fig. 5–13).

Interobserver Variation

Another important consideration is variation between observers. Two examiners often do not derive the same result. The extent to which observers agree or disagree is an important issue, whether we are considering physical examinations, laboratory tests, or other means of assessing human characteristics. We therefore need to be able to express the extent of agreement in quantitative terms.

"This is a second opinion. At first, I thought you had something else."

FIGURE 5–13 One view of a second opinion. (© The New Yorker Collection 1995. Leo Cullum from cartoonbank.com. All rights reserved.)

OVERALL PERCENT AGREEMENT

Table 5–12 shows a schema for examining variation between observers. Two observers were instructed to categorize each test result into one of the following four categories: abnormal, suspect, doubtful, and normal. This diagram might refer, for example, to readings performed by two radiologists. In this diagram, the readings of observer 1 are cross-tabulated against those of observer 2. The number of readings in each cell is denoted by a letter of the alphabet. Thus, A x-rays were read

TABLE 5–12 Observer or Instrument Variation: Percent Agreement

	Reading No. 1			
Reading No. 2	*Abnormal*	*Suspect*	*Doubtful*	*Normal*
Abnormal	[A]	B	C	D
Suspect	E	[F]	G	H
Doubtful	I	J	[K]	L
Normal	M	N	O	[P]

$$\text{Percent agreement} = \frac{A + F + K + P}{\text{Total readings}} \times 100$$

CALCULATING THE OVERALL PERCENT AGREEMENT BETWEEN TWO OBSERVERS

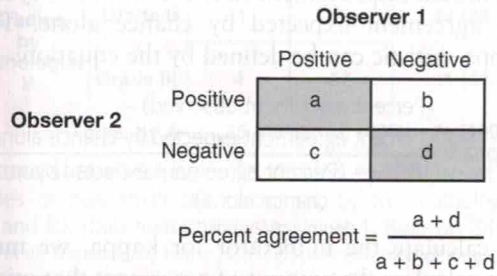


FIGURE 5–14 ▼ Percent agreement when examining paired observations between observer 1 and observer 2.

as abnormal by both radiologists. C x-rays were read as abnormal by radiologist 2 and as doubtful by radiologist 1. M x-rays were read as abnormal by radiologist 1 and as normal by radiologist 2.

As seen in Table 5–12, to calculate the overall percent agreement, we add the numbers in all of the cells in which readings by both radiologists agreed (A + F + K + P), divide that sum by the total number of x-rays read, and multiply the result by 100 to yield a percentage. Figure 5–14 shows the use of this approach for a test with possible readings of either "positive" or "negative."

In general, most persons who are tested have negative results. This is shown in Figure 5–15, in which the size of each cell is drawn in proportion to the number of people in that cell. There is likely to be considerable agreement between the two observers about these negative, or normal, subjects (cell d). Therefore, when percent agreement is calculated for all study subjects, its value may be high only because of the large number of negative findings (cell d) on which the observers agree. Thus, the high value may conceal significant disagreement between the observers in identifying subjects whom at least one observer considers positive.

One approach to this problem, seen in Figure 5–16, is to disregard the subjects who were labeled negative by both observers (cell d) and to calculate percent agreement using as a denominator only the

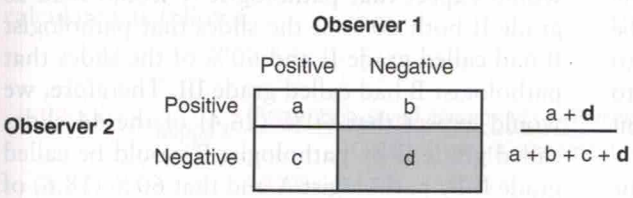


FIGURE 5–15 ▼ Percent agreement when examining paired observations between observer 1 and observer 2, considering that cell d (agreement on the negatives) is very high.

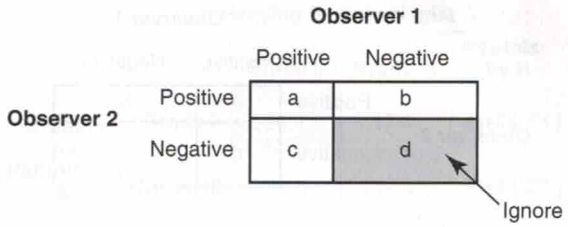


FIGURE 5–16 ▼ Percent agreement when examining paired observations between observer 1 and observer 2, ignoring cell d.

subjects who were labeled abnormal by at least one observer (cells a, b, and c) (Fig. 5–17).

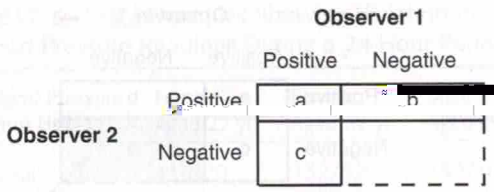
Thus, in the paired observations in which at least one of the findings in each pair was positive, the following equation is applicable:

$$\text{Percent agreement} = \frac{a}{a+b+c} \times 100$$

KAPPA STATISTIC

Percent agreement is also significantly affected by the fact that even if two observers use completely different criteria to identify subjects as positive or negative, we would expect the observers to agree solely as a function of chance.

This can be shown intuitively in the following example: You are the director of a radiology department that is understaffed one day, and a large number of chest x-rays remain to be read. To solve your problem, you go out to the street and ask a few neighborhood residents, who have no background in biology or medicine, to read x-rays as either positive or negative. The first person goes through the pile of x-rays, reading them haphazardly as positive, negative, negative, positive, and so on. The second person does the same, in the same way. Given that both readers have no knowledge, criteria, or standards for reading x-rays, would any of their readings on a specific x-ray agree? The answer is clearly yes; they would agree in some cases, purely by chance.



In the paired observations in which at least one of the observations in each pair was positive, the percent agreement is

$$= \frac{a}{a+b+c} \times 100$$

FIGURE 5–17 ▼ Percent agreement when examining paired observations between observer 1 and observer 2, using only cells a, b and c for the calculation.

However, if we want to know how well two observers read x-rays, we might ask, "To what extent do their readings agree *beyond what we would expect by chance alone?*" In other words, to what extent does the agreement between the two observers exceed the level of agreement that would result just from chance? One approach to answering this question is to calculate the kappa statistic, proposed by Cohen in 1960.[2]

In order to understand kappa, we ask two questions. First, how much better is the agreement between the observers' readings than would be expected by chance alone? This can be calculated as the percent agreement observed minus the percent agreement we would expect by chance alone. This is the numerator of kappa:

(Percent agreement observed)
– (Percent agreement expected by chance alone)

Our second question is, "What is the most that the two observers could have improved their agreement over the agreement that would be expected by chance alone?" Clearly, the maximum that they could agree would be 100% (full agreement—the two observers agree completely). Therefore, the most that we could expect them to be able to improve would be:

100% – (Percent agreement expected by chance alone)

This is the denominator of kappa.

Kappa expresses the extent to which the observed agreement exceeds that which would be expected by chance alone (numerator) relative to the most that the observers could hope to improve their agreement (i.e., 100% – agreement expected by chance alone) [denominator].

Thus kappa quantifies the extent to which the observed agreement that the observers achieved exceeds that which would be expected by chance alone, and expresses it as the proportion of the maximum improvement that could occur beyond the agreement expected by chance alone. The kappa statistic can be defined by the equation:

$$\text{Kappa} = \frac{\text{(Percent agreement observed)} - \text{(Percent agreement expected by chance alone)}}{100\% - \text{(Percent agreement expected by chance alone)}}$$

To calculate the numerator for kappa, we must first calculate the amount of agreement that might be expected on the basis of chance alone. Let us consider data reported on the histologic classification of lung cancer that focused on the reproducibility of the decisions of pathologists in subtyping cases of non–small cell lung carcinoma.[3] Figure 5–18 shows data comparing the findings of two pathologists in subtyping 75 such cases.

The first question is, "What is the observed agreement between the two pathologists?" Figure 5–19 shows the readings by pathologist A along the bottom of the table and those of pathologist B along the right margin. Thus, pathologist A identified 45 (or 60%) of all of the 75 slides as grade II and 30 (or 40%) of the slides as grade III. Pathologist B identified 44 (or 58.7%) of all of the slides as grade II and 31 (or 41.3%) of the slides as grade III. As discussed earlier, the percent agreement is calculated by the following equation:

$$\text{Percent agreement} = \frac{41 + 27}{75} \times 100 = 90.7\%$$

That is, the two pathologists agreed on 90.7% of the readings.

The next question is, "If the two pathologists had used entirely different sets of criteria, how much agreement would have been expected solely on the basis of chance?" Pathologist A read 60% of all 75 slides (45 slides) as being grade II and 40% (30 slides) as grade III. If his or her readings had used criteria independent of those used by pathologist B (e.g., if pathologist A were to read 60% of any group of slides as grade II), we would expect that pathologist A would read as grade II both 60% of the slides that pathologist B had called grade II and 60% of the slides that pathologist B had called grade III. Therefore, we would expect that 60% (26.4) of the 44 slides called grade II by pathologist B would be called grade II by pathologist A and that 60% (18.6) of the 31 slides called grade III by pathologist B

		Grading by Pathologist A: Grade II	Grade III	Totals by B
Grading by Pathologist B	Grade II	41	3	44 (58.7%)
	Grade III	4	27	31 (41.3%)
Totals by A		45 (60%)	30 (40%)	75 (100%)

FIGURE 5–18 ▼ Histologic classification by subtype of 75 slides of non–small cell carcinoma, by two pathologists (A and B). (Data from Ghandur-Mnaymneh L, Raub WA, Sridhar KS, et al: The accuracy of the histological classification of lung carcinoma and its reproducibility: A study of 75 archival cases of adenosquamous carcinoma. Cancer Invest 11:641, 1993.)

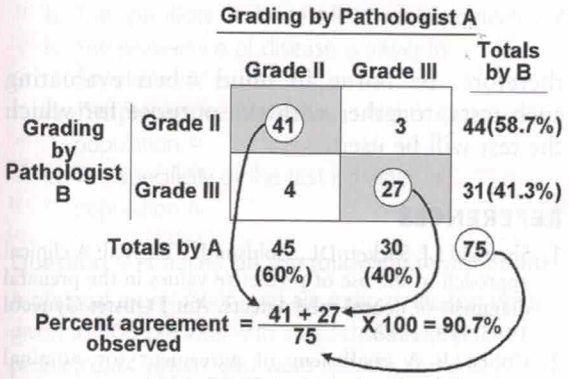


FIGURE 5–19 ▼ Percent agreement by pathologist A and pathologist B. (Data from Ghandur-Mnaymneh L, Raub WA, Sridar KS, et al: The accuracy of the histological classification of lung carcinoma and its reproducibility: A study of 75 archival cases of adenosquamous carcinoma. Cancer Invest 11:641, 1993.)

would also be called grade II by pathologist A (Fig. 5–20). Of the 31 slides called grade III by pathologist B, 40% (12.4) would also be classified as grade III by pathologist A.

Thus, the agreement expected by chance alone would

$$= \frac{26.4}{75} + \frac{12.4}{75} = \frac{38.8}{75} = 51.7\%$$

of all slides read.

Having calculated the figures needed for the numerator and denominator, kappa can now be calculated as follows:

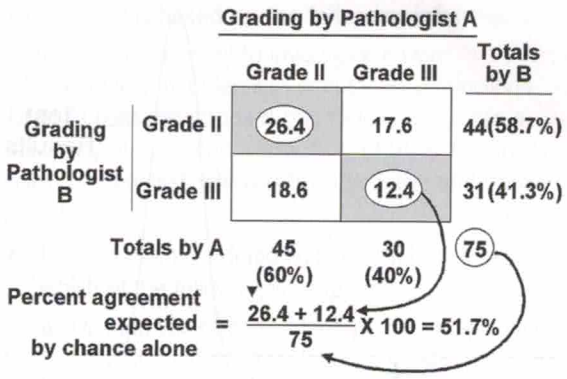


FIGURE 5–20 ▼ Percent agreement by pathologist A and pathologist B expected by chance alone. (Data from Ghandur-Mnaymneh L, Raub WA, Sridar KS, et al: The accuracy of the histological classification of lung carcinoma and its reproducibility: A study of 75 archival cases of adenosquamous carcinoma. Cancer Invest 11:641, 1993.)

Landis and Koch[4] suggested that a kappa greater than 0.75 represents excellent agreement beyond chance, a kappa below 0.40 represents poor agreement, and a kappa of 0.40 to 0.75 represents intermediate to good agreement. Testing for the statistical significance of kappa is described by Fleiss.[5] Considerable discussion has arisen about the appropriate use of kappa, a subject addressed by MacLure and Willett.[6]

Relationship between Validity and Reliability

To conclude this chapter, let us compare validity and reliability using a graphical presentation.

The horizontal line in Figure 5–21 is a scale of values for a given variable, such as blood glucose level, with the true value indicated. The test results obtained are shown by the curve. The curve is narrow, indicating that the results are quite reliable (repeatable); unfortunately, however, they cluster far from the true value, so they are not valid. Figure 5–22 shows a curve that is broad and therefore has low reliability. However, the values obtained cluster around the true value and, thus, are valid. Clearly, what we would like to achieve are results that are both valid and reliable (Fig. 5–23).

$$\text{Kappa} = \frac{\text{(Percent observed agreement)} - \text{(Percent agreement expected by chance alone)}}{100\% - \text{(Percent agreement expected by chance alone)}}$$

$$= \frac{90.7\% - 51.7\%}{100\% - 51.7\%} = \frac{39\%}{48.3\%} = 0.81$$

FIGURE 5–21 ▼ Graph of hypothetical test results that are reliable, but not valid.

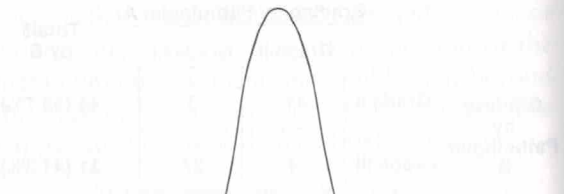


FIGURE 5–23 ▼ Graph of hypothetical test results that are both valid and reliable.

It is important to point out that in Figure 5–22, in which the distribution of the test results is a broad curve centered on the true value, we describe the results as valid. However, the results are valid only for a group (i.e., they tend to cluster around the true value). It is important to remember that what may be valid for a group or a population may not be so for an individual in a clinical setting. When the reliability or repeatability of a test is poor, the validity of the test for a given individual also may be poor. The distinction between group validity and individual validity is therefore important to keep in mind when assessing the quality of diagnostic and screening tests.

Conclusion

This chapter has discussed the validity of diagnostic and screening tests as measured by their sensitivity and specificity, their predictive value, and the reliability or repeatability of these tests. Clearly, regardless of how sensitive and specific a test may be, if its results cannot be replicated, the test is of little use. All these characteristics must, therefore, be borne in mind when evaluating such tests, together with the purpose for which the test will be used.

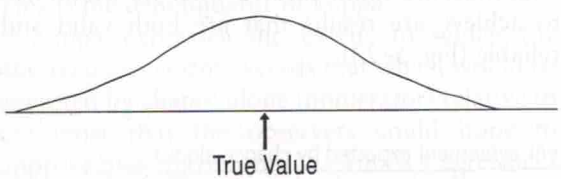


FIGURE 5–22 ▼ Graph of hypothetical test results that are valid, but not reliable.

REFERENCES

1. Sheffield LJ, Sackett DL, Goldsmith CH, et al: A clinical approach to the use of predictive values in the prenatal diagnosis of neural tube defects. Am J Obstet Gynecol 1465:319, 1983.
2. Cohen J: A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37, 1960.
3. Ghandur-Mnaymneh L, Raub WA, Sridhar KS, et al: The accuracy of the histological classification of lung carcinoma and its reproducibility: A study of 75 archival cases of adenosquamous carcinoma. Cancer Invest 11:641, 1993.
4. Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 33:159, 1977.
5. Fleiss JL: Statistical Methods for Rates and Proportions, ed 2. New York, John Wiley & Sons, 1981.
6. Maclure M, Willett WC: Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 126: 161, 1987.

REVIEW QUESTIONS FOR CHAPTER 5

Questions 1, 2, and 3 are based on the information given below:

A physical examination was used to screen for breast cancer in 2,500 women with biopsy-proven adenocarcinoma of the breast and in 5,000 age- and race-matched control women. The results of the physical examination were positive (i.e., a mass was palpated) in 1,800 cases and in 800 control women, all of whom showed no evidence of cancer at biopsy.

1. The sensitivity of the physical examination was: _____
2. The specificity of the physical examination was: _____
3. The positive predictive value of the physical examination was: _____
4. A screening test is used in the same way in two similar populations, but the proportion of false-positive results among those who test positive in population A is lower than that among those who test positive in population B. What is the likely explanation for this finding?
   a. It is impossible to determine what caused the difference
   b. The specificity of the test is lower in population A
   c. The prevalence of disease is lower in population A
   d. The prevalence of disease is higher in population A
   e. The specificity of the test is higher in population A

Question 5 is based on the following information:

A physical examination and an audiometric test were given to 500 persons with suspected hearing problems, of whom 300 were actually found to have them. The results of the examinations were as follows:

PHYSICAL EXAMINATION

	Hearing Problems	
Result	*Present*	*Absent*
Positive	240	40
Negative	60	160

AUDIOMETRIC TEST

	Hearing Problems	
Result	*Present*	*Absent*
Positive	270	60
Negative	30	140

5. Compared with the physical examination, the audiometric test is:
   a. Equally sensitive and specific
   b. Less sensitive and less specific
   c. Less sensitive and more specific
   d. More sensitive and less specific
   e. More sensitive and more specific

Question 6 is based on the following information:

Two pediatricians want to investigate a new laboratory test that identifies streptococcal infections. Dr. Kidd uses the standard culture test, which has a sensitivity of 90% and a specificity of 96%. Dr. Childs uses the new test, which is 96% sensitive and 96% specific.

6. If 200 patients undergo culture with both tests, which of the following is correct?
   a. Dr. Kidd will correctly identify more people with streptococcal infection than Dr. Childs
   b. Dr. Kidd will correctly identify fewer people with streptococcal infection than Dr. Childs
   c. Dr. Kidd will correctly identify more people without streptococcal infection than Dr. Childs
   d. The prevalence of streptococcal infection is needed to determine which pediatrician will correctly identify the larger number of people with the disease

Questions 7 and 8 are based on the following information:

A colon cancer screening study is being conducted in Nottingham, England. Individuals 50 to 75 years old will be screened with the Hemoccult test. In this test, a stool sample is tested for the presence of blood.

7. The Hemoccult test has a sensitivity of 70% and a specificity of 75%. If Nottingham has a prevalence of 12/1,000 for colon cancer, what is the positive predictive value of the test?
8. If the Hemoccult test result is negative, no further testing is done. If the Hemoccult test result is positive, the individual will have a second stool sample tested with the Hemoccult II test. If this second sample also tests positive for blood, the individual will be referred for more extensive evaluation. What is the effect on net sensitivity and net specificity of this method of screening?
   a. Net sensitivity and net specificity are both increased
   b. Net sensitivity is decreased and net specificity is increased
   c. Net sensitivity remains the same and net specificity is increased
   d. Net sensitivity is increased and net specificity is decreased
   e. The effect on net sensitivity and net specificity cannot be determined from the data

Questions 9 to 12 are based on the information given below:

Two physicians were asked to classify 100 chest x-rays as abnormal or normal independently. The comparison of their classification is shown in the following table:

CLASSIFICATION OF CHEST X-RAYS BY PHYSICIAN 1 COMPARED WITH PHYSICIAN 2

	Physician 2		
Physician 1	*Abnormal*	*Normal*	Total
Abnormal	40	20	60
Normal	10	30	40
Total	50	50	100

9. The simple, overall percent agreement between the two physicians out of the total is: 70%
10. The overall percent agreement between the two physicians, removing the x-rays that both physicians classified as normal, is: 57.1%
11. The value of kappa is: 0.40
12. This kappa represents which kind of agreement? b
    a. Excellent
    b. Intermediate to good
    c. Poor

chap

6

At this point
and screenin
sick and hea
identified as
"How can w
the disease in
cation is imp
necessary to
establish prio
health progr
questions ab
cation is im
natural histo
available, th
compared w
them. Furthe
are available
or medical t
surgical proc
pare the eff
therapy. The
we need a q
prognosis in

This chap
which progn
terms for a g
history of di
chapter; late
to intervene
improve prog
randomized
appropriate
18 discusses
earlier point
maximize th

To discus
schematic re
of disease in

Point A r
Often, this