Clinical biostatistics

XXXI. On the sensitivity, specificity, and discrimination of diagnostic tests

Alvan R. Feinstein, M.D.* *West Haven, Conn.*
The Cooperative Studies Program Support Center and the Department of Medicine of the West Haven Veterans Administration Hospital, and the Departments of Medicine and Epidemiology of the Yale University School of Medicine

In 1947, Yerushalmy[18] introduced the terms *sensitivity* and *specificity* as statistical indexes of the efficiency of a diagnostic test. The sensitivity of the test would indicate its capacity for making a correct diagnosis in confirmed positive cases of the disease. The specificity would indicate the capacity for correct diagnosis in confirmed negative cases.

These concepts need not be restricted to diagnostic tests alone and can be applied to a variety of tests used for identifying clinical conditions. The relationship between clinical conditions and the results of tests is commonly shown in the following "fourfold" table:

RESULT OF TEST	CONFIRMED CONDITION	
	Positive	*Negative*
Positive	True positive	False positive
Negative	False negative	True negative

In this table, the column headings denote the patient's confirmed condition; the row headings denote the result of the test; and the four interior "cells" denote whether the patient has been correctly or falsely diagnosed as either positive or negative. According to Yerushalmy's delineation, *sensitivity* would be the number of true positive cases divided by the total number of confirmed positive cases, which is the sum of true positive plus false negative cases. *Specificity* would be the number of true negative cases divided by the total number of confirmed negative cases, which is the sum of true negative plus false positive cases.

Yerushalmy introduced these terms while performing studies of observer variability among radiologists; and his work was an important contribution to the scientific growth of epidemiology. He helped shatter some of the complacency with which unverified diagnostic statements have been accepted and tabulated in epidemiologic statistics, and his indexes of sensitivity and specificity provided a quantitative method for expressing the problems. The indexes have now become widely applied as statistical tools in the analysis of clinical epidemiologic data. The phrase *sensitivity and specificity*, like the phrase *range of normal*, is now an established part of the mathematical concepts that constantly appear in medical statistics.

Like the statistical concepts associated with *range of normal*[8], however, the conventional mathematical ideas associated with *sensitivity* and *specificity* are inadequate for the real world activities of clinical medicine. One of the main problems is in temporal direction. A clinician wants to use the test predictively; the epidemiologic indexes are often constructed in the wrong chronologic direction, emanating from a "backward" rather than "forward" viewpoint. A second problem arises from oversimplification. Not all diagnoses can be dichotomously cited as either *yes* or *no;* among the other categories to be considered are *probably yes, prob-*

Supported by Public Health Service Grant No. HS 00408 from the National Center for Health Services Research and Development.

*Professor of Medicine and Epidemiology, Yale University School of Medicine, New Haven, Conn.; Senior Biostatistician, Cooperative Studies Program Support Center, Veterans Administration Hospital, West Haven, Conn.

ably no, and *uncertain.* A third problem is caused by clinical imprecision in the idea of a "diagnostic test". Some diagnostic tests are used to detect the existence of a particular disease, whereas others are used to confirm it. The statistical expression of the test's "ability" will be inadequate unless the purpose of the test is suitably considered. The greatest problem of all, however, arises from neither the unsatisfactory direction of the arithmetic nor the clinical naiveté of the mathematics. The problem is caused by inadequate choices of "control" groups. The patients whose conditions are tested during the evaluation procedures are seldom selected in a way that will determine the true discrimination of the test.

A. Temporal direction

1. Choice of symbols. To discuss chronologic and certain other issues in the mathematics, we need to have some symbols for the four groups of people who appeared in the "cells" of the foregoing table. The choice of these symbols creates a problem of its own. There may be a great deal of observer variability among physicians practicing medicine, but there is even more "symbol variability" among the investigators who discuss sensitivity and specificity.

In Yerushalmy's original paper, he used no symbols. Subsequent authors[1, 3, 11, 13–17] have chosen diverse arrangements of two-letter or one-letter expressions to provide a magnificently creative array of such symbols as α, β, a, b, R, K, ρ, $f_1(\rho)$, $f_2(\rho)$, p, 1 − p, 1 − b, ϵ, η, ρ^+, ρ^-, p_{ij}, cf, Se, In, and P[R | Y]. In the absence of any national or international efforts at standardization of this statistical babel, an author newly entering the field is free to choose whatever symbols he wishes. I shall use the a, b, c, d symbols that are reasonably familiar to most clinicians looking at the contents of a fourfold table.

The numbers of people present in the four cells of the table cited earlier will thus be listed as:

RESULT OF TEST	CONFIRMED CONDITION	
	Positive	*Negative*
Positive	a	b
Negative	c	d

I shall also use s to represent the sensitivity of the test and f to represent the specificity. With this convention, the sensitivity of the test is

$$s = \frac{a}{a + c},$$

and the specificity of the test is

$$f = \frac{d}{b + d}.$$

2. Predictive use of a test. At the time a test is first evaluated, its proponents usually assemble a population of people whose condition was known to be positive or negative. When the test was performed in these people and when the numerical frequencies of the data were arranged in the fourfold pattern of the table, the investigators calculated the indexes of sensitivity and specificity. If the results seemed sufficiently encouraging, the test might be accepted into general clinical usage. The troubles would then begin.

The original investigators started with a population whose condition had already been confirmed, but the clinician who later uses the test starts with patients whose condition is unknown. The purpose of the test is to predict (or identify) what the patient's condition really is. In receiving the result of the test, an investigator therefore wants to know its predictive accuracy, not its sensitivity or specificity. He wants to know how well the test would perform for an unknown patient, not its capacity in people who really did not need the test because their correct diagnosis had already been established. If the test result is positive, is the patient's actual condition likely to be positive? If the result is negative, is the actual condition likely to be negative?

To answer these two questions, we need a different set of indexes. In the sensitivity-specificity calculations, the denominators were chosen on the basis of what was found after the diagnoses had already been confirmed. For clinically useful indexes, we would want the denominators to depend on the predictions made by the test. We would therefore want to have an index of positive accuracy—denoting how often the test was correct when its result was positive. We would also want an index of

negative accuracy for the correctness of negative results.

If we denote positive accuracy by v, and negative accuracy by g, these indexes would be expressed respectively as

$$v = \frac{a}{a + b} \text{ and } g = \frac{d}{c + d}.$$

An alternative complementary approach is to consider the "false positive rate", $1 - v$, which is the number of false positives divided by the total number of positive results in the test; and the "false negative rate", $1 - g$, which is the number of false negatives divided by the total number of negative results.

With either set of approaches, the denominators now contain the sums of either the positive results or the negative results of the test. We have avoided the denominator "criss-cross" that occurs when negative and positive results are combined for calculating sensitivity and specificity. The positive accuracy and the false positive rates of the test are based on the total of true positive and false positive results; the negative accuracy and the false negative rates are based on the total of true and false negatives.

To illustrate the predictions with some numbers, let us assume that the original investigator assembled 50 patients known to be positive and 100 patients who were confirmed as negative. After the test was performed, the results were as follows:

RESULT OF TEST	CONFIRMED CONDITION	
	Positive	*Negative*
Positive	45	10
Negative	5	90

For these data, the sensitivity of the test is $45/(45 + 5) = 90\%$; and the specificity of the test is $90/(90 + 10) = 90/100 = 90\%$. In its predictive value, however, the test has a positive accuracy of $45/(45 + 10) = 45/55 = 82\%$; and a negative accuracy of $90/(90 + 5) = 90/95 = 95\%$.

The false positive and false negative rates would give these data an even more meaningful clinical expression. In the cited example, the false positive rate is 18% (= 100% − 82%); and the false negative rate is 5% (= 100% − 95%). These distinctions demonstrate that a test with equally high rates of sensitivity and specificity can give substantially different rates of false positive and false negative results when it is applied in clinical practice.

3. Effect of population ratios. We can now get ready for another dismaying discovery. Suppose the original investigator had gotten different numbers of patients in the two groups he assembled for the evaluation. Suppose he had applied the test to 20 patients who were known to be positive and to 200 patients who were the known negative "controls". Since the sensitivity and specificity of the test are presumably its inherent properties, they would have remained intact at 90% each. After the test was performed, the fourfold table of results would show the following:

RESULT OF TEST	CONFIRMED CONDITION	
	Positive	*Negative*
Positive	18	20
Negative	2	180

In these data, the sensitivity of the test (= 18/20) and the specificity (= 180/200) are each 90%. A dramatic change has occurred, however, in the test's predictive value. The positive accuracy rate is only 47% (= 18/38), so that a false positive result occurred in more than half the people who were diagnosed by the test. On the other hand, the negative accuracy has improved to 99% (= 180/182) so that only 1% of the patients were false negatives.

We now recognize not only that the sensitivity and specificity of a diagnostic test will fail to indicate its predictive value, but also that the predictive value will depend entirely on the ratio of confirmed positive and confirmed negative people to whom the test was applied. This ratio is best noted by contemplating the prevalence rate of the confirmed positive condition. This prevalence rate equals the number of confirmed positive cases divided by the total number of cases under study.

For readers who are willing to endure some algebra, the relationships can be symbolically shown as follows. Let N be the total number of people for whom the test is evaluated. Let P be the prevalence rate (or proportion) of people whose confirmed condition is positive. [$P = (a + c)/N$]. The ingredients and totals of our fourfold table

would then be numerically expressed as follows:

People with confirmed condition positive	= PN
People with true positive results	= sPN
People with false negative results	= (1 − s)PN
People with confirmed condition negative	= (1 − P)N
People with true negative results	= f(1 − P)N
People with false positive results	= (1 − f)(1 − P)N

The rate of positive accuracy of the test would be

$$\frac{sPN}{sPN + (1 - f)(1 - P)N} = \frac{sP}{sP + (1 - f)(1 - P)}$$

The rate of negative accuracy of the test would be

$$\frac{f(1 - P)N}{f(1 - P)N + (1 - s)PN} = \frac{f(1 - P)}{f(1 - P) + (1 - s)P}$$

In the first numerical example, we had 50 confirmed positive cases and 100 negative controls, so that P = 50/(100 + 50) = 1/3. With s = f = 0.9, we then have a positive accuracy of (0.9)(.33)/[(0.9)(.33) + (0.1)(.67)] = .30/[.30 + .07] = .30/.37 = 82%. The negative accuracy would be (.9)(.67)/[(.9)(.67) + (.1)(.33)] = .60/[.60 + .03] = .60/.63 = 95%.

In the second numerical example, we had 20 confirmed positive cases and 100 negative controls, so that P = 20/[20 + 200] = .90. With s = f = 0.9, the positive accuracy is (.9)(.09)/[(.9)(.09) + (.1)(.91)] = .081/[.081 + .091] =.081/.172 = 47%. The negative accuracy is (.9)(.91)/[(.9)(.91) + (.1)(0.9)] = .819/[.819 + .009] = .819/.828 = 99%.

By looking at the algebraic symbols for these expressions, we can note that P appears in the numerator of the expression for positive accuracy. If P is large (i.e., a value close to 1), the tested population will contain a preponderance of confirmed positive patients and very few negative "controls". With this large value of P, the value of 1 − P will be small (i.e., a value close to 0), and so the positive accuracy of the test will approach a value of 1 (i.e., 100%). Conversely, since 1 − P appears in the numerator of the expression for negative accuracy, this expression will take on its highest value when P is very small (approaching 0), so that 1 − P is essentially equal to 1.

By altering the prevalence rate of confirmed positive cases in the tested population, an investigator can thus make the results show almost any values that he wishes to achieve for positive or negative accuracy, regardless of whatever be the sensitivity and specificity of the test. For example, suppose the test is not much better than tossing a coin, having a sensitivity and specificity each equal to 50%. Let us choose 100 positive people and 10 negative "controls" for evaluation. The results will show 50 true positives, 50 false negatives, 5 false positives and 5 true negatives. The test will therefore have a good "batting average" for positive accuracy [= 50/(50 + 5) = 50/55 = 91%] but a bad one for negative accuracy [= 5/(50 + 5) = 9%]. On the other hand, if we chose our evaluation group to contain 10 confirmed positive people and 100 negative "controls", these batting averages would be exactly reversed, with a 9% positive accuracy and a 91% negative accuracy.

If the test is being used to diagnose a disease, the value of P will indicate the prevalence of the disease in the tested population. With a very high prevalence of the disease, a test with a sensitivity and specificity that are no better than tossing a coin might thus have excellent predictive accuracy. If the disease has a very low prevalence in the tested population, the same test will have high accuracy for the negative prediction of "excluding" the disease.

A different type of distress will occur when a test of apparently high sensitivity and specificity is removed from its evaluation in a hospital population and is applied for diagnostic screening in a general population. Suppose an investigator reports a sensitivity of 0.95 and a specificity of 0.85 in a new diagnostic test for cancer. When we apply this test in screening, we can expect the prevalence of cancer to be about 150 per 100,000 patients, so that P = .0015. By substituting in the previous formula, we can promptly determine the rate of positive accuracy of the test. Since s = 0.95, f = 0.85, and P = .0015, the rate of positive accuracy will be (.95)(.0015)/[(.95)(.0015) + (.15)(.9985)] = (.00143)/[.00143 + .14978] = .00143/.15120 = .00942 = 0.9%. Thus when the test gives a positive result, the likelihood will be less than 1% that the patient actually has cancer!

The problems caused by these differences in the prevalence of a tested condition have been

thoroughly discussed by Vecchio[17] and by Sunderman and Van Soestbergen.[16] Both of these authors' presentations contain tables showing the extraordinarily wide range of values that can occur for "false positive" and "false negative" results of a diagnostic test whose sensitivity and specificity are evaluated in populations with different prevalence rates of the disease.

4. Alternative pattern of "sampling". In the foregoing procedures, the investigator who evaluated the test did his "sampling" from the confirmed cases. He began by choosing one group of people who were known to be positive for the disease (or target condition) and a "control" group of people who were known to be negative. Since these groups were selected by the investigator, he could determine their size and would thereby set the prevalence rate of the disease.

An alternative method of getting the tested population is for the investigator to choose the groups according to the results of the test. He would select one group of people with positive results in the test and another group with negative results. He would then determine the actual conditions in the two groups, and calculate the rate of false positives and false negatives. In this circumstance, the size of the two groups chosen by the investigator would determine the rate of a positive result in the tested population, and we would encounter a different arrangement of the algebraic phenomena.

Let R = the proportion of people with a positive test result, so that R = (a + b)/N. Assuming that the test has a fixed sensitivity and specificity, and letting J = s + f − 1, we can go through an array of mathematical manipulations (which I shall spare the reader here) to show the following result:

$$\text{Rate of positive accuracy of test} = \frac{s}{J}\left[1 + \frac{f-1}{R}\right]$$

$$\text{Rate of negative accuracy of test} = \frac{f}{J}\left[1 - \frac{1-s}{1-R}\right]$$

To check this calculation, consider the second numerical example cited earlier, where s = f = 0.9 and where we had 38 positive test results and 182 negative test results, so that R = 38/(38 + 182) = 38/220 = 0.173. Substituting into the formula just cited, we find that the positive accuracy of the test is [.9/.8][1 − (.1/.173)] = [1.125][.422] = .474 = 47%. The negative accuracy of the test is [.9/.8][1 − (.1/.827)] = [1.125][.8791] = .989 = 99%. These results are the same as what we obtained before with the formulas based on values of P, rather than R.

The two formulas just listed will indicate how the rates of positive and negative accuracy can be affected by the choice of R. The algebra is more complex than the arrangements noted earlier for P, but a scan of the associated values of s, f, and J will usually suggest an appropriate choice of R to create suitable effects on the rates of positive or negative accuracy.

Thus, regardless of whether the investigator gets his evaluation groups by choosing cases from the confirmed condition of the patients or from the results of the test, he can arbitrarily alter the calculated rates for false positive and false negative values. Sensitivity and specificity therefore indicate properties that are unaffected by numerical caprice in the respective sizes of the groups used to evaluate a test. This statistical constancy is the desirable feature that has made these two indexes achieve such widespread acceptance. As we shall see later, however, sensitivity and specificity depend on much more than numerical ratios alone. The basic issue is the clinical composition of the tested groups, not just their sizes.

B. Summary indexes

The next major problem occurs as a result of the statistical penchant for "data reduction". Rather than having two different expressions, such as *sensitivity* and *specificity* or *false positive rate* and *false negative rate,* we might prefer to combine the two expressions into a single index. Several statistical indexes of association can be used to create a single value that "summarizes" the results found in a fourfold table. The available indexes include[10] such splendid algebraic eponyms as Guttman's λ, Yule's Q, Yule's Y, Pearson's C, and Tschuprow's T, as well as the ∅ coefficient and the coefficient of tetrachoric correlation.

Two other indexes, however, have become particularly popular for summarizing the results of a fourfold table dealing with the sensitivity and specificity of a diagnostic test. One of these is called an index of "validity". It is determined as (a + d)/N, and it represents the total

Table I. *Palpation vs. thermometry for measuring temperature**

Temperature estimated by palpation	*Actual temperature*			*Total*
	≥ 39° C (major fever)	*38-38.9° C (minor fever)*	*No fever*	
≥ 39° C *(major fever)*	15	3	3	21
38-38.9° C *(minor fever)*	19	43	15	77
No fever	3	55	993	1051
Total	37	101	1011	1149

*Table rearranged from data presented by Bergeson and Steinfeld.[2]

number of correct predictions divided by the total number of predictions. From our previous symbols, we can recall that a = sPN and d = f(1 − P)N. The index of validity will therefore be the sum of sP and f(1 − P). The constituents of this sum clearly indicate how the "validity" of the test can be altered by the way that the investigator chooses P, the prevalence of the positive condition. To get a high validity score, the investigator need merely choose a high or low value of P according to the relative magnitudes of s and f.

Another summary index, introduced by W. J. Youden[19] (and sometimes called "Youden's J"), has an algebraic structure that ultimately becomes equal to the sum of sensitivity plus specificity minus 1. Since this index has the advantage of being unaffected by the choice of P, Youden's J is a preferred way of combining sensitivity and specificity into a single value.

No matter how a summary index is contrived, however, it will suffer from two major disadvantages:

1. By combining everything into a single value, we lose track of whether the diagnostic test is better in sensitivity or specificity. For example, if Youden's J has a value of 0.55, we would have no idea of whether the sensitivity is 0.95 and specificity is 0.60; or vice versa.

2. More importantly, in using any of the fourfold summary indexes, we accept the idea that the results of the evaluation procedure can readily be listed in a fourfold table. According to this idea, the presence of the disease can be expressed as a simple *yes* or *no,* and the results of the diagnostic test can also be expressed in the same dichotomy. This double-dichotomy arrangement creates a gross and often erroneous oversimplification of the realities of clinical diagnosis. In many circumstances, the disease cannot be cited as definitely present or definitely absent, and the diagnostic test may yield the result of *maybe* (or *uncertain*) rather than *yes* or *no.* If both the presence of the disease and the results of the test are cited in a 3-category rating scale, however, the calculations of sensitivity and specificity become much more complicated; and the summary index must deal with a ninefold rather than fourfold table.

To avoid these complexities, the data analyst may try to compress a table with nine or more cells into one that contains only four cells. For this compression, the data analyst gets to draw two arbitrary lines that determine the dichotomous "break points" for the consolidations that form the rows and columns of the new table. The arbitrary choice of these lines can strikingly alter what happens to the sensitivity and specificity of the test.

To illustrate the problem, let us consider Table I, which contains data taken from a recent report in which Bergeson and Steinfeld[2], working in a Child Care Clinic at the Johns Hopkins Hospital, tried to determine whether the fever discerned with a thermometer could be detected equally well by a nurse's palpation of the child's forehead or chest. The ninefold arrangement in Table I shows the bivariate frequencies of the data obtained with each method of examination, using a 3-category scale for reporting the result as **no fever, minor fever** (38-38.9° C) and **major fever** (≥ 39° C). The investigators decided to categorize the results of palpation according to three designations: *correct, too high,* and *too low.* Correct results, shown in the three downward diagonal cells of the table, occurred in 15 + 43 + 993

= 1051 cases. In 21 cases (= 15 + 3 + 3), the palpation method gave a falsely high result; and in 77 cases (= 55 + 3 + 19), palpation yielded a falsely low result.

The investigators laudably made no effort to calculate a dichotomous sensitivity and specificity for the "palpation test". In many other similar circumstances, however, such calculations would have been performed with one of at least three different ways of compressing the Bergeson-Steinfeld data. One approach would be to draw perpendicular dichotomous lines at fever vs. no fever. With this approach the numbers in Table I would become:

	ACTUAL CONDITION	
RESULT OF PALPATION	*Fever*	*No fever*
Fever	80	18
No fever	58	993

The second arrangement would be to draw the dichotomous lines at major fever vs. nonmajor fever. With this arrangement, Table I would become:

	ACTUAL CONDITION	
RESULT OF PALPATION	*Major fever*	*Not major fever*
Major fever	15	6
Not major fever	22	1106

A third arrangement depends on a previous decision about clinical tactics. Let us decide that we will always use a thermometer to take the patient's temperature if palpation indicates the intermediate condition of **minor fever**. Furthermore, for purposes of using palpation as a "screening test", let us assume that we are not really interested in circumstances where the thermometer shows only **minor fever**. What we really want to know is the reliability of palpation in "screening" for major fever. With these assumptions, five cells are removed from the original nine-fold table, and it reduces to the following four-fold table:

	ACTUAL CONDITION	
RESULT OF PALPATION	*Major fever*	*No fever*
Major fever	15	3
No fever	3	993

We can now calculate the sensitivity and specificity of palpation, as shown in three different arrangements of the same basic set of data.

In the first arrangement, s = 58% (= 80/138) and f = 98% (= 993/1011). The false positive and false negative rates are 18% and 6%, respectively. In the second arrangement, s = 41% (= 15/37) and f = 99% (= 1106/1112). The respective false positive and false negative rates are 29% and 2%. In the third arrangement, s = 83% (= 15/18) and f = 99.7% (= 993/996). The false positive rate is 17%; and the false negative, 0.3%. We can thus get three different sets of values for sensitivity and specificity, or for false positive and negative rates, according to the way we decide to dichotomize the data. If the original table had contained more categories in both directions—so that the results were arranged in a 4 × 4 or even larger pattern of cells—the opportunities for disagreement would be even greater when the data were doubly dichotomized.

In addition to this difficulty, a separate problem that arises in the construction of any "two-way" contingency table—no matter how many cells it contains—is the assumption that the entity being evaluated is the univariate result of a single test. Many medical diagnoses depend on an aggregate of the results found in several different variables, not just in one. For example, in acute myocardial infarction, the clinical diagnosis would depend on certain combinations of symptoms, electrocardiographic data, and laboratory tests. In acute rheumatic fever, the Jones diagnostic criteria call for an enumerated collection of entries from certain "major" and "minor" manifestations. A test procedure based on input from just one variable is obviously inadequate for determining the sensitivity and specificity of these complex diagnoses. We would need to use an expression that contains multivariate constituents. An example of such a variable would be *fulfillment of composite criteria for diagnosis of acute myocardial infarction*. The categories of this variable could be expressed in terms such as **yes** or **no** (or **uncertain**).

This method of citing the result of a multivariate diagnostic procedure would allow us to use a 2-way table for comparing the enumerated

data of whatever method was employed to confirm the patients' correct diagnoses. On the other hand, because the constituent multivariate elements are lost in a single expression such as **yes** or **no**, we would have no direct way of determining the causes of erroneous results when they occur. To track down the sources of false positive and false negative diagnostic errors, we would have to go back and start with each of the multivariate constituents.

C. Relationship of index and purpose

Both of the statistical difficulties that have just been mentioned could be overcome (or at least reduced) with a more sophisticated set of mathematical indexes for expressing the relationships. Instead of using Youden's J, or the "index of validity", or any other indexes that depend on doubly dichotomous data in a fourfold table, we could use indexes of association that allow the variables to have polytomous (more than two) categories. Such indexes would include Kendall's tau, Goodman and Krushkal's G, Cicchetti's statistic[4], and some of the various "kappa" statistics described by Fleiss[9] or the "lambda" statistics described by Hartwig[12]. (If worst came to worst, or perhaps to best, we could simply enumerate the results according to the proportions that were too high, correct, and too low). To consider the correlation between multivariate constituents of data and the patient's confirmed condition, we could use some of the diverse correlation coefficients that can be derived from multiple linear regression or discriminant function analysis.

These statistical improvements in managing multicategory or multivariate data, however, will not solve a more fundamental problem in describing the effectiveness of a test. What seems to be almost wholly overlooked in clinicostatistical strategies for calculating a test's effectiveness is the purpose for which the test is used.

1. The three types of diagnostic test. Diagnostic tests are employed for at least three different purposes: discovery, confirmation, and exclusion. During various types of "screening" procedures, we use a *discovery test*. Examining people who seem healthy, with no clinical complaints to suggest the presence of a particular disease, we often search for that disease in a clinically "silent" form. Examples of such discovery tests in lanthanic patients are the uses of a serum calcium measurement for hyperparathyroidism, a fasting blood sugar for diabetes mellitus, or a rectal examination for rectal cancer.

A *confirmation test* is employed in situations where we have strong suspicions that the disease is present. The purpose of the test is to verify this suspicion. The performance of bronchoscopy with microscopic examination of biopsy tissue is a confirmation test for lung cancer; and a glucose tolerance test provides confirmation for diabetes mellitus.

An *exclusion test* is usually employed to "rule out" the presence of a disease when it is suspected. Such a test is usually too expensive or inconvenient to be employed merely for discovery purposes during routine "screening". For example, a stool guaiac examination might be used for the screening discovery of colonic cancer, but a more elaborate roentgenographic or colonoscopic examination would be needed to "rule out" the disease if its presence is suspected. Certain exclusion tests are cheap enough and convenient enough to be used for screening purposes. Thus, when an appropriate skin test for tuberculosis is negative, the presence of active disease can usually be excluded, although a positive test will neither discover nor confirm active tuberculosis.

Some tests are good for only one of these three purposes. Some are good for two. Some can be used for all three. For example, the performance of sigmoidoscopy, together with biopsy and histologic examination when appropriate, can generally be used to discover, confirm, and exclude cancer of the rectum. A glucose tolerance test can be used to confirm and to exclude the presence of diabetes mellitus, but is generally too inconvenient for purposes of screening discovery. The histologic examination of tissue from a bronchoscopic biopsy is an excellent way to confirm lung cancer, but cannot be used to exclude the disease or to discover it during routine screening.

Since diagnostic tests are employed for these different purposes, the statistical indexes of efficiency should be arranged accordingly.

2. Requirements of detection and confirmation. In a discovery test, we want reasonably high sensitivity. If the disease is present, it

should be found, even at the risk of getting a high rate of false positive results. [We are willing to take this risk because a discovery test, when positive, is usually followed by a confirmation test]. In an exclusion test, we want the sensitivity to be even higher than in a discovery test. Unless the sensitivity is 1 or close to 1, the risk of a false negative result would keep us from being confident that a negative test has excluded the disease.

The discovery and exclusion tests are thus both intended to have a high sensitivity for detecting the disease when it is present. To get the particularly high sensitivity that is sought in an exclusion test, we must be willing to pay the appropriate clinical price. Thus, to test urine for sugar is a good, cheap, convenient way of "screening" for the discovery of diabetes mellitus, but the urine test will regularly give some false negative results. To measure fasting blood sugar is a more expensive and less convenient discovery procedure, but it is more diagnostically effective because it has a lower false-negative rate than the urine test. If we want to rule out diabetes mellitus with certainty, however, we cannot rely on either of these procedures. We would have to use the much more expensive and cumbersome mechanism of the glucose tolerance test, which, in this instance, would be both an exclusion and a confirmation test.

By contrast, in a confirmation test, we want extremely high specificity, with few or no false positive results. If the test shows that the disease is present, we want to be sure that it is present. We would have no real objection to occasional false negative results, since the confirmation test will probably be ordered after an exclusion test was used to find any cases that might otherwise be missed as false negatives.

3. ***Combinations of tests.*** A single test can seldom be excellent for the goals of both detection and confirmation. With rare exception, the same procedure cannot be sensitive enough to find all cases of the disease while simultaneously being specific enough to avoid false positive identifications. For example, the chest X-ray is a quite sensitive but non-specific way of finding lung cancer. Almost all patients with lung cancer have abnormal roentgenograms, but not all people with positive roentgenograms turn out to have lung cancer. Conversely, a positive bronchoscopic biopsy is a quite specific but non-sensitive way of identifying lung cancer. The bronchoscopic biopsy almost never gives false positive results, but it regularly will miss lung cancers that are located at inaccessible sites.

For these reasons, many diagnostic tests are regularly used in tandem. A high sensitivity test is used to find the disease; and a positive result is followed by a high specificity test that will confirm the diagnosis by "excluding" its possible falsehood. Because of these tandem arrangements, the best statistical appraisal of the results will depend on a suitable arrangement of the paired tests. In such an arrangement, the result of the pair might be called *negative* if the detection test is negative; and the paired result would be called *positive* only if both the detection test and the confirmation test are positive. The positive and negative results of this kind of paired arrangement would have both high specificity and high sensitivity.

D. Choice of the tested populations

There are important clinical reasons for trying to solve some of the problems that have just been discussed. Perhaps the most important reason is that this form of correlation between the result of a test and the patient's actual condition is the best way of making clinical sense out of the statistical chaos that now exists in demarcating the "range of normal"[8]. If "normality" is determined purely on a univariate basis, according to arbitrary statistical boundaries for a distribution of data, the demarcation will indicate the zone of customary values for the test, but not their clinical connotations in health or disease. If the demarcated zone is to have these clinical connotations, the demarcations must be established in direct correlation with an actual condition of health or disease. This type of correlation can be achieved and evaluated only through the type of bivariate arrangements we have been discussing.

The discussion so far has been concerned, however, only with the defects of existing clinico-statistical strategies and with ways of improving the defects. Unfortunately, these mathematical improvements will not solve the really fundamental biostatistical problems of

diagnostic tests. Like so many other sophisticated statistical procedures, the complex indexes of association produce elegant but superficial algebra. The indexes can provide useful methods of quantitative expression for what has been observed—but the calculations are totally dependent on what is submitted as the observed data. And the fundamental biostatistical problem lies in the choice of the populations that are the sources of the data.

1. The role of clinical suspicion. If we are going to use a test for different diagnostic purposes, it must be evaluated in groups of people who suitably represent the different diagnostic challenges. These people cannot be chosen merely according to whether or not they were demonstrated to have the disease in question. Since the preceding clinical suspicions will affect the choice of a test and the evaluation of the test's performance, the tested population must at least be divided according to the existence of clinical suspicions. We would thus choose one group of people who constitute the ordinarily healthy population for whom the test would be used, during "screening", as a detection test. The second group of people would have medical conditions that aroused our suspicion of the disease and that made us want to confirm it or exclude it.

The customary fourfold diagnostic table would thus be converted into the following "eightfold" table:

RESULTS OF TEST	ACTUAL CONDITION	
	Positive	*Negative*
Screened population:		
Positive	a′	b′
Negative	c′	d′
Suspected population:		
Positive	a″	b″
Negative	c″	d″

If these populations are going to approximate reality, we would want P, the prevalence of the actual disease, to be low in the screened population and high in the suspected population.

When the test results are correlated with the patients' actual condition, we would calculate at least two sets of values for sensitivity and specificity—one set for the screened population and another set for the suspected population. Thus, instead of a single value for sensitivity [which would be $(a' + a'')/(a' + c' + a'' + c'')$], we would determine two separate values: $a'/(a' + c')$ for the screened population and $a''/(a'' + c'')$ in the suspected population. Two analogous calculations would be done for specificity, using the respective b and d values in the screened and suspected populations.

2. The role of pathologic derangement. By inspecting this eightfold arrangement of data, we can begin to see why a particular test might have not one set of values for sensitivity and specificity, but several different sets. Suppose a positive result in the test depends on the disease having produced a certain level of pathologic derangement. When this level of derangement occurs, the diseased persons almost always develop symptoms that arouse suspicions of the disease. In such suspected patients, the test will therefore have high sensitivity. On the other hand, if the disease is present without having reached the prerequisite level of pathologic derangement, the patient may be asymptomatic and part of a screened population. In such a population, the diagnostic test may have low sensitivity.

Once we begin to contemplate a pathologic derangement[7], rather than the particular entity that is called a "disease", we can also recognize the causes of many false positive or disproportionately positive results that can destroy the value of a diagnostic test. For example, suppose the positive result of a particular diagnostic test really depends on a derangement in the patient's nutritional status, but suppose we want to employ this test for the diagnosis of cancer. For the evaluated population, we choose the diseased group from hospitalized patients with cancer, and the non-diseased group from healthy technicians, secretaries, and other staff personnel. Since patients whose cancer is severe enough to require hospitalization are often malnourished, the results of their test are usually positive. Since the staff personnel are well nourished, their test results are negative. We emerge from the evaluation process with the belief that we have found an excellent new diagnostic test for cancer: the sensitivity and specificity values are quite high.

After the test begins to be applied, we may be chagrined to discover that it really has low

sensitivity and low specificity. The test fails to detect the neoplasms of asymptomatic well-nourished patients with cancer; and it gives false positive diagnoses of cancer for malnourished patients with stroke, chronic cardiopulmonary disease, or certain enteropathies. Because we failed to include such patients in the original test population, we did not discover the inefficiency of the test until after it became clinically popular.

3. Surrogate vs. pathognomonic tests. The term *pathognomonic* is usually applied to a clinical manifestation that uniquely indicates a particular condition. For example, the palpation of spontaneous movement within a suitable sized suprapubic mass in a woman would be pathognomonic of pregnancy. This term can also be used for paraclinical procedures that either delineate, demonstrate, or otherwise identify a particular disease. For example, the histologic findings in an appropriate tissue specimen will be pathognomonic of cancer or hepatitis; a specified set of values in a glucose tolerance test will be pathognomonic of diabetes mellitus.

In a *surrogate* test, we examine an entity that will be used to represent or approximate the disease we want to identify. Examples of surrogate tests are pap smears for cancer, serum glutamic oxalic transaminase (SGOT) for hepatitis, chest X-ray for tuberculosis, electrocardiogram for myocardial infarction, or urine sugar for diabetes mellitus.

A pathognomonic test is seldom evaluated for sensitivity and specificity. We may worry about observer variability when a pathologist interprets a tissue specimen; or about the standards of glucose ingestion, specimen timing, and chemical measurement when a laboratory performs a glucose tolerance test; but we are not concerned that the test itself may be misleading.

It is the surrogate tests that create the main problems of sensitivity and specificity. A surrogate test does not identify the disease; it identifies something else that we hope will denote the disease. We often use surrogate tests because they are simpler, cheaper, and more convenient than the corresponding pathognomonic test. The surrogate test may also be more sensitive. For example, a measurement of serum alkaline phosphatase may detect metastatic cancer that has been missed by a liver biopsy; and a positive chest X-ray can detect tuberculosis that has not shown tubercle bacilli in the microscopic examination of sputum. To compensate for these advantages, surrogate tests often produce false results and the problem of evaluating sensitivity and specificity.

Because the procedure is a surrogate test, it depends on a pathologic entity that is different from the one we are trying to diagnose. To contemplate sources of false positive and false negative results, we must therefore contemplate the mechanisms that might "trigger" a test into errors of omission or commission. These mechanisms will consist of alternative pathologic derangements or clinical conditions. Thus, inflammation may create a false positive pap smear for cancer; and inaccessibility of the desquamated cells may make the pap smear falsely negative. The electrocardiogram may fail to show a myocardial infarction if taken too early after the acute attack and may give false positive results because of some other myocardopathy. Many chemical tests give falsely high results in response to alternative diseases and drugs; and the results can be falsely lowered under other appropriate clinical conditions.

4. The process of discrimination. For all these reasons, a proper evaluation of the surrogate procedures that are called *diagnostic tests* would require them to receive several different challenges in discrimination. The test must be able to discriminate among a variety of pathologic derangements that might simulate either the target disease or an entity in the clinical and paraclinical spectrum of that disease. The various groups of patients who enter the evaluated population must be selected according to their suitability for providing these challenges. If patients are chosen merely because they do or do not have the target disease, the discrimination of the test will not be adequately evaluated.

The choice of patients to provide appropriate challenges will depend on both the medical spectrum of the disease and its diagnostic co-morbidity. The medical spectrum[5] of the disease refers to the array of clinical and paraclinical laboratory abnormalities that it can produce. The diagnostic co-morbidity of the disease consists of other diseases that might

be mistaken for it. Diagnostically co-morbid diseases are usually entities occurring in the same topographic location of the body or producing somewhat similar morphologic or other paraclinical abnormalities.

For example, the medical spectrum of primary lung cancer would indicate patients with hemoptysis, with major weight loss, and with abnormal chest roentgenograms. The spectrum of diagnostic co-morbidity for lung cancer would include patients with non-neoplastic pulmonary diseases (such as tuberculosis and chronic bronchitis) and with metastatically neoplastic pulmonary lesions. To evaluate the discrimination of a proposed new test for lung cancer, we would therefore want to challenge the test with patients who represent different parts of the medical and co-morbid spectrum.

Our investigated population might thus include the following groups of people: asymptomatic patients with lung cancer; patients with lung cancer and only primary symptoms, such as hemoptysis; patients whose lung cancer symptoms include such systemic effects as major weight loss; patients whose lung cancer manifestations include such metastatic effects as hepatomegaly or bone pain; asymptomatic patients with other causes of pulmonary disease; hemoptytic patients with other pulmonary disease; patients with major weight loss due to other diseases; and patients with hepatomegaly or bone pain due to other diseases.

In a more general statement of principles, the populations used to evaluate the discrimination of a diagnostic test for Disease X should consist of representatives from the following groups of people:

1. Patients with Disease X who are asymptomatic.

2. Patients with Disease X who are symptomatic with a diverse collection of manifestations that cover the medical spectrum of the disease.

3. Patients without Disease X who have other diseases that have produced overt manifestations similar to those in the medical spectrum noted in Group 2.

4. Patients without Disease X who have other diseases that can mimic Disease X's pathologic derangement by occurring in a similar location or by having similar paraclinical dysfunctions.

The sensitivity of a test used for discovery purposes in "screening" will depend on its capacity to identify patients in Group 1. The test's sensitivity for exclusion purposes will depend on its performance in identifying members of groups 1 and 2. The specificity of the test will depend on its avoidance of false positive results in groups 3 and 4. These four groups would seem to be a minimum demarcation of the necessary comparisons, but additional subgroupings would be needed in appropriate circumstances.

The complexity of these arrangements may seem distressing, but they are ultimately less distressing than the continued proliferation of diagnostic tests whose inadequacies escape initial evaluations because the initial evaluations did not contain suitable challenges. The oversimplification of the existing tactics for getting "control" groups and calculating statistical indexes has led to the spawning of many tests that are grossly unsatisfactory for clinical purposes.

To deal with clinical reality requires a confrontation with clinical complexity. The new arrangements proposed here are both feasible and analyzable after the appropriate data have been assembled. The performance of such complex analyses is not at all a novel idea. It has been, in fact, performed for many years during a generally unquantified procedure called *clinical judgment*[5]. With increasing advances in technology, clinicians will increasingly have to evaluate the costs, risks, and diagnostic discrimination of new diagnostic tests. If these evaluations are to provide sensible clinical science, the subtleties and complexities of clinical judgment must be acknowledged, adapted, and incorporated into the plans for choosing the patients who are tested and for quantitatively expressing the results.

• • •

Author's note: Many people have asked me how I can find time to prepare these essays every two months. The answer is that I can no longer do so. For 1975, the essays will appear in this journal at quarterly rather than bimonthly intervals, in the issues of January, April, July, and October.—A. R. F.

References

1. Bennett, B. M.: On comparisons of sensitivity, specificity and predictive value of a number of diagnostic procedures, Biometrics **28:**793-800, 1972.
2. Bergeson, P. S., and Steinfeld, H. J.: How dependable is palpation as a screening method for fever? Clinical Pediatr. (Phila.) **13:**350-351, 1974.
3. Berkson, J.: "Cost-utility" as a measure of the efficiency of a test, J. Amer. Stat. Assn. **47:**246-255, 1947.
4. Cicchetti, D. V.: A new measure of agreement between rank ordered variables, Proc. Am. Psychol. Assoc. **7:**17-18, 1972.
5. Feinstein, A. R.: Clinical judgment (reprinted edition), Huntington, N. Y., 1974, Robert E. Krieger Publishing Co.
6. Feinstein, A. R.: The pre-therapeutic classification of co-morbidity in chronic disease, J. Chronic Dis. **23:**455-469, 1970.
7. Feinstein, A. R.: An analysis of diagnostic reasoning. I. The domains and disorders of clinical macrobiology, Yale J. Biol. Med. **46:**212-232, 1973.
8. Feinstein, A. R.: Clinical biostatistics. XXVII. The derangements of the range of normal, CLIN. PHARMACOL. THER. **15:**528-540, 1974.
9. Fleiss, J. L.: Statistical methods for rates and proportions, New York, 1973, John Wiley & Sons, Inc.
10. Freeman, L. C.: Elementary applied statistics: For students in behavioral science, New York, 1965, John Wiley & Sons, Inc.
11. Greenhouse, S. W., and Mantel, N.: The evaluation of diagnostic tests, Biometrics **6:**399-412, 1950.
12. Hartwig, F.: Statistical significance of the Lambda coefficients, Behav. Sci. **18:**307-310, 1973.
13. Mantel, N.: Evaluation of a class of diagnostic tests, Biometrics **7:**240-246, 1951.
14. Muic, V., Petres, J. J., and Telisman, Z.: Validity of a diagnostic test designated by a single function, Methods Inf. Med. **12:**244-248, 1973.
15. Nissen-Meyer, S.: Evaluation of screening tests in medical diagnosis, Biometrics **20:**730-755, 1964.
16. Sunderman, F. W., and Van Soestbergen, A. A.: Laboratory suggestions: Probability computations for clinical interpretations of screening tests, Am. J. Clin. Pathol. **55:**105-111, 1971.
17. Vecchio, T. J.: Predictive value of a single diagnostic test in unselected populations, N. Engl. J. Med. **274:**1171-1173, 1966.
18. Yerushalmy, J.: Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques, Pub. Health Rep. **62:**1432-1449, 1947.
19. Youden, W. J.: Index for rating diagnostic tests, Cancer **3:**32-35, 1950.