INDEX FOR RATING DIAGNOSTIC TESTS

W. J. YOUDEN, PH.D.

THE results of studies of a diagnostic test are customarily tabulated as follows:

	Classified by Test		
Individuals	*Positive*	*Negative*	*Total*
Known diseased	a	b False negatives	a+b
Healthy or control	c False positives	d	c+d

When choosing between two or more diagnostic tests, the decision has to be based upon such sets of data, taking into consideration the practicability of the several procedures.

The ideal diagnostic test should discriminate unerringly between diseased and healthy individuals. When such a test is available, there is no statistical problem. The search for an unerring test yields a series of tests that achieve partial success. There is, therefore, a need for an index to rate diagnostic tests in an objective manner and to provide a means of deciding whether two diagnostic tests really differ in their capacity to discriminate between healthy and diseased individuals. Diagnostic tests undergo modifications in attempts to improve their performance. The medical researcher needs a statistical tool to assist him in detecting as early as possible whether a particular modification has led to an improvement in the results obtained with the diagnostic test.

There are two questions that arise in appraising a diagnostic test to which it is not possible to give completely satisfactory answers. First, whenever a diagnostic test is subjected to study, the question arises as to how sure the experimenter can be that his controls are really healthy and that all his diseased patients actually have the disease. The latter part of the question is usually met by using advanced cases in which the disease has been demonstrated to be present. The controls are often simply young, available, and assumed-to-be-healthy individuals. The unsuspected presence of the disease in a control is probably not as serious as the contrast between the two groups in other respects, chiefly, of course, in age. If the diagnostic test is in some way responsive to age, then the results may be due to age differences and not to the absence or presence of the disease. There is no statistical protection against this confounding of causes other than to ensure that the control group resembles the diseased group as closely as possible in all respects save the presence of the disease.

The other question has to do with the relative seriousness of the two types of error a diagnostic test may make: false positives and false negatives. There is usually an instinctive reaction against a test that is subject to false negatives, since presumably these individuals are lost if treatment is deferred until too late. On the other hand, false positives exact a price, quite aside from the unpleasant shock to the individuals concerned. If, as a result of false positives, the *limited facilities* for treatment are expended upon individuals who do not need treatment, then the efficiency of the use of the facilities is lowered. Put another way, a test that never gives a false positive means that available facilities are used at maximum efficiency and therefore with best over-all consequences to the population as a whole. It is, in fact, not a statistical matter to decide what weights should be attached to these two types of diagnostic error.

The purpose of an index of performance is to reduce a table of data, like that just given, into one figure that will adequately characterize the diagnostic test. An index, apparently not previously employed, can be derived by the argument given in the next paragraph.

From the National Bureau of Standards, Washington, D.C.

Received for publication, November 10, 1949.

The proportion of diseased individuals correctly classified is a/a+b. It seems appropriate to charge against this the proportion, b/a+b, incorrectly classified, leaving $\frac{a-b}{a+b}$ as a measure of the success of the test on the diseased group. As further evidence that this is a reasonable measure, note that, if a = b, or in other words if the test is equally likely to report a diseased individual negative as positive, it has no discriminative power on the diseased group. A similar argument for the control group gives $\frac{d-c}{c+d}$ as a measure of success with that group. Let the average of these two be taken as the index. This assumes false positives to be as undesirable as false negatives.

$$J=\tfrac{1}{2}\left[\frac{a-b}{a+b}+\frac{d-c}{c+d}\right]=\frac{a}{a+b}+\frac{d}{c+d}-1$$
$$=\frac{ad-bc}{(a+b)(c+d)}$$

The index, J, is seen to be also equal to the sum, diminished by unity, of the two fractions showing the proportions *correctly* diagnosed for the diseased and control groups. The expression may also be written as a fraction in which the numerator is made up of the product of the numbers correctly diagnosed diminished by the product of the numbers incorrectly classified. The denominator is the product of the totals in the diseased and control groups.

This index has certain desirable features.*

1. The possible range of values for the index is from zero to one inclusive. (It is expected that the test will show a greater proportion of positive results for the diseased group than for the control.)

2. The index has the value zero whenever a diagnostic test gives the same proportion of positives for both diseased and control groups regardless of what that proportion is. Such a test is obviously worthless.

Group	*Diagnosis*		*Total*
	+	−	
+	40	10	50
−	8	2	10

$$J=\frac{(40)(2)-(10)(8)}{(50)(10)}=\frac{80-80}{500}=0$$

3. The index becomes unity only when *both* false positives and false negatives are not present. If only one type of error is made, the index is controlled by that error.

Group	*Diagnosis*		*Total*
	+	−	
+	50	0	50
−	10	40	50

$$J=\frac{(50)(40)-(10)(0)}{(50)(50)}=\frac{40}{50}=0.80$$

4. The index is independent of the relative sizes of the control and diseased groups.

Group	*Diag.*		*Total*
	+	−	
+	90	10	100
−	10	40	50

$$J=\frac{3600-100}{5000}=0.70$$

Group	*Diag.*		*Total*
	+	−	
+	9	1	10
−	10	40	50

$$J=\frac{360-10}{500}=0.70$$

5. The index is independent of the absolute sizes of the control and diseased groups.

Group	*Diag.*		*Total*
	+	−	
+	23	2	25
−	1	9	10

$$J=\frac{207-2}{250}=0.82$$

Group	*Diag.*		*Total*
	+	−	
+	230	20	250
−	10	90	100

$$J=\frac{20700-200}{25000}=0.82$$

6. All tests that have the same index make the same total number of misclassifications per hundred patients (figured separately for diseased and control groups).

* These include those listed by M. G. Kendall: The Advanced Theory of Statistics. London. Charles Griffin and Co., Ltd. 1947; Vol. I, p. 310.

Group	Diag. +	Diag. −	Total	Index	Misclassifications per 100+	Misclassifications per 100−	Total
+	50	0	50	$\frac{1750}{2500}=0.70$	0	30	30
−	15	35	50				
+	80	20	100	$\frac{7200-200}{10000}=0.70$	20	10	30
−	10	90	100				
+	1	0	1	$\frac{70}{100}=0.70$	0	30	30
−	30	70	100				

7. It is possible to calculate a standard error for the index. Naturally, the larger the groups, the more reliable, in the sense of being free from random sampling variation, is the experimental value of the index. Any index is incomplete unless there is also available some means of setting up confidence limits for the estimate of the index.

The index, when written in the form

$$J=\frac{a}{a+b}+\frac{d}{c+d}-1$$

is seen to be, apart from the constant, the sum of two fractions that refer to the proportions correctly diagnosed for the diseased and control groups. These proportions are properties of a diagnostic test.

If $\frac{A}{A+B}$ and $\frac{D}{C+D}$ represent the true proportions, then the standard errors of estimates of the proportions are respectively $\sqrt{\frac{AB}{(A+B)^3}}$ and $\sqrt{\frac{CD}{(C+D)^3}}$. (The standard error of an observed proportion, p, binomially distributed, is $\sqrt{P(1-P)/N}$). The standard error of their sum, and consequently of the index "J" is

$$S.E._J=\sqrt{\frac{AB}{(A+B)^3}+\frac{CD}{(C+D)^3}}.$$

In practice, the data available from the study of a test are used as an estimate of the true proportions. So long as the numbers in the groups are not too small (20 or more), and the index not close to zero or one, the binomial distribution approximates the normal distribution and the usual procedures are employed for setting up confidence limits.

The following data* were obtained with a proposed diagnostic test for cancer and with a modified version of the test.

ORIGINAL TEST				MODIFIED TEST			
Group	Diag. +	Diag. −	Total	Group	Diag. +	Diag. −	Total
+	95	6	101	+	40	11	51
−	75	33	108	−	7	23	30
	$J=\frac{3135-450}{10908}=0.246$				$J=\frac{920-77}{1530}=0.551$		
	$S.E._J=\sqrt{\frac{570}{1030301}+\frac{2475}{1259712}}$				$S.E._J=\sqrt{\frac{440}{132651}+\frac{161}{27000}}$		
	$=\sqrt{.000553+.001965}$				$=\sqrt{.003317+.005963}$		
	$=\sqrt{.002518}=0.050$				$=\sqrt{.009280}=0.0963$		
	95% confidence interval 0.148 − 0.344				0.362 − 0.740		

To obtain 95 per cent confidence intervals for the indexes twice (more accurately 1.96), the standard error is added and subtracted from the index. The intervals do not overlap, so that we may conclude that the improvement in the index shown by the modified test is a real effect.

The proper way to compare two indexes is by means of the *t* test.

$$t=\frac{\text{difference between indexes}}{\text{S.E. of their difference}}=\frac{0.551-0.246}{0.109}=\frac{0.305}{0.109}=2.80$$

where

$$S.E._{diff}=\sqrt{(S.E.)_1^2+(S.E.)_2^2}=\sqrt{.002518+.009280}$$
$$=\sqrt{.011798}=0.109$$

The 5 per cent value for *t* is 1.96, the 1 per cent value is 2.58. Since the value of *t* for these indexes is 2.80, the difference between

*Courtesy of Dr. F. Homburger, Tufts College Medical School. See Evaluation of Diagnostic Tests for Cancer. III. pp. 15-25, this issue of CANCER.

them is said to be significant at the 1 per cent level, meaning that there is less than one chance in a hundred that a difference as large as this would have been found if the two tests were in reality equally effective.

The formula given for the standard error of the index is not suitable for use with small numbers, say when there are fewer than 20 in a group. With small numbers, the standard error will be very large, and special statistical procedures are required to compare two diagnostic tests.

Summary

In summary, the paper proposes a new index for measuring the performance of diagnostic tests. Diagnostic tests have the task of correctly designating which are the diseased individuals in a population. These tests may err in giving negative tests for diseased individuals (false negatives) and in giving positive tests for healthy individuals (false positives). The formula for the index of performance is

$$J = \frac{ad - bc}{(a+b)(c+d)}$$

where, of (a+b) diseased patients, "a" are correctly diagnosed and "b" are false negatives, and where of (c+d) controls, "d" are correctly reported and "c" are false positives. The index has the value zero if the test reports the same proportion of positive tests for both control and diseased groups. It has the value unity when, and only when, there are neither false positives nor false negatives resulting from the test. The difference between two diagnostic tests may be evaluated by means of the standard errors of the indexes.