The Evaluation of Tests

S. W. Martin*

ABSTRACT

In order to correctly evaluate a test, at least four attributes should be measured: namely, sensitivity, specificity, accuracy and precision. Sensitivity is the proportion of diseased animals which are correctly identified, whereas specificity is the proportion of healthy animals which are correcty identified. These two attributes are important, not only because of the reasons implied by their definition but because they influence both the apparent prevalence of disease and the proportion of test-positive animals which are actually diseased.

The ability of a test to give a true measurement of the substance being measured, its accuracy, and its ability to give consistent results on the same sample, its precision, are good measures of quality control. Both these attributes influence the sensitivity and specificity of the test. Inaccuracies and inconsistencies arise from the test itself, the technician and the nature of the sample being tested.

RÉSUMÉ

Pour évaluer correctement une épreuve, il faut tenir compte d'au moins quatre qualités: la sensibilité, la spécificité, la fidélité et la précision. La sensibilité représente la proportion d'animaux malades qu'une épreuve permet d'identifier correctement, tandis que la spécificité représente la proportion d'animaux sains qu'elle permet aussi d'identifier avec exactitude. Ces deux qualités tirent leur importance non seulement des raisons inhérentes à leur définition, mais aussi du fait qu'elles influencent tant la prédominance apparente d'une maladie que la proportion des animaux réacteurs qui sont réellement malades.

L'habilité d'une épreuve à mesurer exactement une substance donnée, sa fidélité, son habilité à donner des résultatts constants avec un même échantillon et sa précision, constituent des bons critères de sa valeur. Ces qualités influencent la sensibilité et la spécificité d'une épreuve. Des inexactitudes et des inconsistances originent d'une épreuve elle-même, du technicien et de la nature de l'échantillon en cause.

INTRODUCTION

A variety of tests, any method or process used to ascertain the nature of something, are frequently employed to detect abnormal states of health in animals. Such tests are aids in separating a large group of animals into two groups, i.e. those most likely to be diseased and those most likely to be healthy (12), in estimating the prevalence and/or incidence of disease in populations (14) and in formulating a diagnosis in individual animals (5).

Regardless of the reason(s) for using a test, every test has potential errors and these errors must be identified and measured if the results of a test are to be interpreted wisely (13). Recognizing this, most workers have attempted to assess the efficacy of their test(s). However, different words have been used to describe the "errors" which they claim to have measured (10). Although standard methods for the evaluation of tests have been published (3, 5, 7, 18, 20) they do not appear in the literature frequently used by veterinarians. Thus, this manuscript discusses the attributes of a test which need to be measured and proposes a standard nomenclature to describe these attributes.

Sensitivity and Specificity of a Test

Sensitivity and specificity are two important attributes of a test. Sensitivity is the ability of a test to correctly detect an animal with a specified disease. This defini-

*Epidemiology Section, Department of Veterinary Microbiology and Immunology, University of Guelph, Guelph, Ontario N1G 2W1.

Submitted February 6, 1976.

tion of sensitivity must be differentiated from its common usage to describe the ability of a test to detect small amounts of antigen or antibody (1). Specificity is the ability of a test to correctly detect healthy animals and/or those not having the specified disease. Lack of sensitivity leads to false negative results and lack of specificity leads to false positive results.

These two attributes are described further in Table I. In Table I a population or random sample of animals is categorized with respect to health status and test results. To simplify the description a specified disease is considered as present (D+) or absent (D–) and the test results as positive (T+) or negative (T–). Animals with a specified disease are called "diseased". Animals without that disease are called "healthy", although they may have diseases other than the one specified. From Table I, sensitivity is the proportion of diseased animals which are test positive, i.e. the proportion a/(a+c) and specificity is the proportion of healthy animals which are test negative, i.e. the proportion d/(b+d). These may be expressed as conditional probabilities (6) with sensitivity represented by p(T+/D+) and specificity by p(T–/D–). In addition, the false negative rate is p(T–/D+) and the false positive rate is p(T+/D–). When expressed as probabilities or proportions, sensitivity and the false negative rate sum to one [p(T+/D+) + p(T–/D+) = 1] and specificity plus the false positive rate sum to one [p(T–/D–) + p(T+/D–) = 1].

In addition to the ability of a test to detect diseased animals, i.e. its sensitivity, it is of value to know the proportion of test positive animals which are diseased. This proportion p(D+/T+) = a/(a+b) is known as the predictive value (8, 17, 19) of a positive test result. The conditional probabilities, p(T+/D+) and p(D+/T+), representing the sensitivity and predictive value respectively, are rarely equal. The predictive value is a function of sensitivity p(T+/D+), the false positive rate (p(T+/D–)) and the prevalence of disease (p(D+)). This fact is easily shown using Bayes' Theorem (15) as follows:

$$p(D+/T+) = p(T+/D+) \times p(D+) / [p(T+/D+) \times p(D+) + p(T+/D-) \times p(D-)]$$

where p(D–) = 1 – p(D+) (i.e. the proportion of healthy animals is found by subtracting the proportion diseased from 1). In general, it is simpler to calculate the predictive value using the results of a 2 x 2 table with a format similar to Table I to determine the proportion a/(a+b).

The following examples are given to clarify the concepts just described. Assume that it is necessary to separate a group of cattle into two subgroups: namely, one which is most likely to have disease Y and another which is most likely healthy or does not have disease Y. The screening test will be a standard white blood cell (WBC) count with a screening level of 13,000 WBC's/mm^3. Calves with a WBC count equal to or greater than 13,000 will be termed diseased (having disease Y), whereas those with a lesser count will be deemed healthy (not having disease Y). The actual health status will be determined by a complete "diagnostic work-up". The screening level of 13,000 cells provides a sensitivity of 80% and a specificity of 93%. The results anticipated following the use of the WBC count in a group of calves of which 10% are diseased are given in Table II and the anticipated results when the prevalence of disease is 1% are given in Table III.

In both instances (Tables II and III) the test correctly detects 80% of diseased animals and 93% of healthy animals. When the prevalence of disease is 10%, the apparent prevalence is 14.3% and the predictive value is 56%. When the prevalence of disease is 1%, the apparent prevalence is 7.7% and the predictive value is 10.4%.

These results demonstrate a general rule: namely, that when the disease is rare, only a very small proportion of test positive animals are in fact diseased (8, 15, 17, 19). Further, the apparent prevalence (p(T+)) is often a very poor estimate of the actual prevalence (p(D+)) of the disease. The true prevalence of a disease can only be estimated if the error rates, (p(T–/D+) and p(T+/D–)), of the test(s) are known. In this case, a formula to calculate the true prevalence of disease is:

$$p(D+) = p(T+) - p(T+/D-) / [1-(p(T+/D-)+p(T-/D+))] \quad (10)$$

Figures 1 and 2 provide more general examples of the phenomenon just described. Figure 1 portrays the effect of the false positive rate on the apparent prevalence of disease. Only the test with a specificity of 99.995% provides a reasonably accurate indication of the level of disease when the true prevalence is below 0.07% (7 per 10,000). Figure 2 demonstrates that even

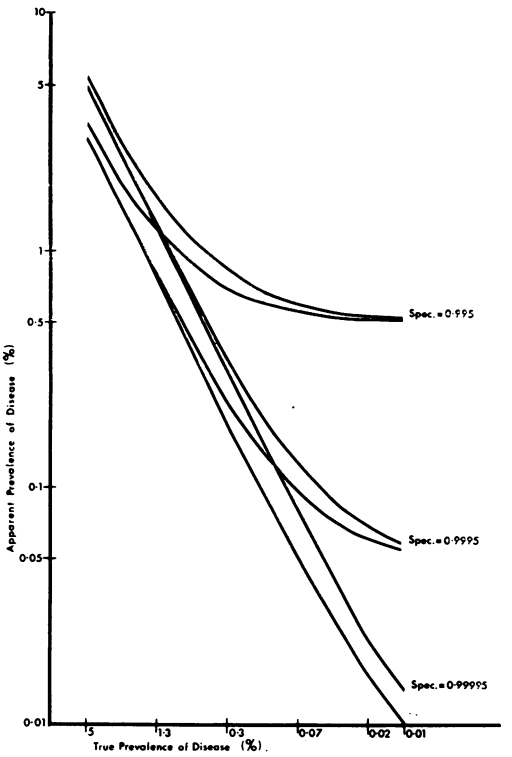

Fig. 1. The relationship between apparent prevalence and true prevalence at different levels of sensitivity and specificity. (The width of the band for each level of specificity represents the difference between a sensitivity of 62% — lower boundary and 100% — upper boundary.)

with a specificity of 99.995%, only approximately 60% of test positive animals are actually diseased.

In order to determine the sensitivity and specificity of a test, clearly defined "diseased" and "healthy" groups are needed (18). In addition, the methods used to define the health status should be independent, in a biological sense, of the test under investigation. That is, if the test is a serological technique then the health status should be determined by a nonserological method, such as microbial culture. The methods for defining health status do not need to be 100% effective, although they should approach this level.

In general, the sensitivity of a serological test can be determined by noting the proportion of positive results in those animals which are positive on microbial cultures (12). However, often one cannot assess the specificity of the test by noting the proportion of negative results among those animals which were negative on microbial cultures. This practice will usually result in gross underestimates of the actual speci-

TABLE I. Classification of a Population of Animals with Respect to Health Status and Test Results

Test Results	Health Status: Have Specified Disease (D+)	Health Status: Do Not Have Specified Disease (D−)	Total
Positive (T +)	a	b	a + b
Negative (T −)	c	d	c + d
Total	a + c	b + d	a+b+c+d = N

The letters a, b, c and d represent an arbitrary number of animals in each test result + health status category.

Sensitivity	= a/(a+c)	= p(T+/D+)
Specificity	= d/(b+d)	= p(T−/D−)
Prevalence	= (a+c/N	= p(D+)
Apparent prevalence	= (a+b)/N	= p(T+)
Predictive value	= a/(a+b)	= p(D+/T+)

TABLE II. The Expected Results of a Test for the Detection of Disease Y in Calves. Sensitivity of Test is 80%, Specificity is 93% and the Prevalence of Disease is 10%

Test Results	Health Status: Diseased (D+)	Health Status: Healthy (D−)	Total
Positive (T+) . .	80	63	143
Negatives (T−).	20	837	857
Total.	100	900	1000

Apparent prevalence	= 143/1000	= 14.3%
Predictive value	= 80/143	= 55.9%

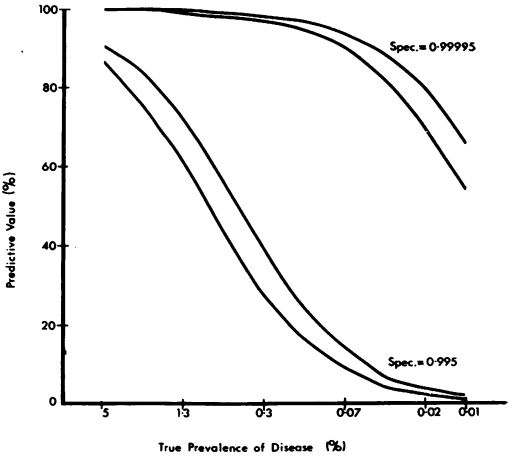

Fig. 2. The relationship between predictive value and disease prevalence at different levels of sensitivity and specificity. (The width of the band for each level of specificity represents the difference between a sensitivity of 62% — lower boundary and 100% — upper boundary.)

TABLE III. The Expected Results of a Test for the Detection of Disease Y in Calves. Sensitivity of Test is 80%, Specificity is 93% and the Prevalence of Disease is 1%

	Health Status		
Test Results	Diseased (D+)	Healthy (D−)	Total Total
Positive (T+) . .	8	69	77
Negative (T−) .	2	921	923
Total.	10	990	1000
Apparent prevalence	= 77/1000	= 7.7%	
Predictive value	= 8/77	= 10.4%	

ficity (12), primarily because many of the animals which were negative on culture are, in fact, infected. To obviate this problem, some workers conduct the test in a population of animals which are known to be not infected (9). Another method is to compare the proportion of test negative animals among those negative to at least two other tests (4). This method will be discussed in more detail under the topic of "series testing".

Often, in an attempt to establish sensitivity and specificity, the results of one test are compared to those of another test. This does not establish sensitivity or specificity, but only *relative* sensitivity and *relative* specificity (18). Such comparisons should be made only if the sensitivity and specificity of the standard test are known and if they approach 100%. Otherwise, such comparisons may lead to false conclusions and delay the identification of tests which are superior to the standard test (6, 20).

Two additional comments on sensitivity and specificity are in order. First, for any given test, sensitivity and specificity usually are inversely related. Thus if the screening level is altered to increase the sensitivity, then specificity decreases and vice versa. This is a consequence of an overlapping distribution of the parameter being measured (in the present example, the distribution of WBC counts) between the animals with a specified disease and those without that disease. For example, if the distribution of WBC counts in cattle is normal (gaussian), with the healthy animals having average counts of 10,000 ± 2000 and the diseased cattle having counts of 18,000 ± 6000, then a screening level of 12,948 cells detects 80% of all diseased cattle and 93% of all healthy cattle. This is easily verified using statistical methods similar to those used to assess type I and type II errors (15). (The assumption of "normality" of distribution for the parameter simplifies the mathematics involved in demonstrating the effect of different screening levels on both sensitivity and specificity. It is possible, or even quite likely, that the distribution of the parameter is nonnormal in one or the other or both groups of animals. However, the general inverse relationship between sensitivity and specificity is still present under these conditions (7)). The initial screening level of 13,000 WBC's may be decreased with a resultant increase in sensitivity and decrease in specificity (Table IV). If the screening level were raised, sensitivity would decrease and specificity increase. This phenomenon is true of all tests if there is an overlapping distribution of the parameter in the diseased and healthy groups, regardless of the test's sensitivity and specificity.

A second consideration is the stability of a test's sensitivity and specificity from one population to another. If the relationship between the parameter being measured and the disease is stable between populations, then the sensitivity should also be stable. However, often the standardization of the test itself is not consistent between populations and in this case the sensitivity may vary considerably (1).

Specificity is usually not stable from one population to another (7). There are a number of reasons for this variation, including the presence of other factors (cross-reacting organisms or other diseases) which alter the distribution of the parameter being measured in the "healthy" segment of the population. This nonstability makes it necessary to ensure adequate standardization as well as equal sensitivity and specificity levels before comparing test results from two or more populations.

TABLE IV. The Relationship[a] Among Screening Level, Sensitivity and Specificity of a White Blood Cell Count

Screening Level (Cells/mm³)	Sensitivity (%)	Specificity (%)
12,948.	80	93
11,784.	85	87
10,308.	90	56

[a]Assumes normal distribution of white blood cells with $\bar{X}_{(D-)}$ = 10,000 ± 2,000 (SD) and $\bar{X}_{(D+)}$ = 18,000 ± 6,000

The Use of More Than One Test

In some cases the use of one test may not be adequate to fulfill the needs of a testing program and thus two or more tests might be required. One method of using two tests is to initially test all animals with one test, often the test with a high sensitivity and then to repeat the test on the positive samples or animals with a second test, usually with the more specific test. This method is quite useful where a sensitive inexpensive test is available for use on all animals and a sensitive and more specific test (not as inexpensive or convenient) is available for limited use. An example of this method is the use of the rose bengal plate test (RBPT) in screening all cattle for brucellosis. Any sera reacting to the RBPT are tested with the complement fixation test (2).

Another test method is to apply two or more tests to all samples or animals (18). In this case, the overall sensitivity and specificity may be altered, depending on how the test results are interpreted. When two tests are used, "parallel" interpretation denotes all samples or animals which are positive to one or the other or both tests as positive. This tends to increase the sensitivity and decrease the specificity of the combined tests. "Series" interpretation denotes only those samples or animals which are positive to both tests as positive. This tends to increase specificity and decrease the sensitivity of the combined tests. There are very few instances where the use of multiple tests will increase both sensitivity and specificity (13).

Series testing is a useful method of assessing *relative* sensitivity and specificity in situations where it is too difficult or costly to assess the true health status (18). The results of the test under consideration are compared with the results of a battery of other tests. Those samples or animals which gave positive reactions to all of the tests in the battery are designated "diseased" and those giving negative reactions to all of the tests in the battery are designated "healthy". The number of "diseased" and "healthy" samples or animals serve as denominators for the relative sensitivity and relative specificity calculations. In such comparisons, the term "relative" is important, since one is not assessing the actual sensitivity or specificity but rather the degree of agreement among tests. It is noteworthy that the percentage agreement may be very high when the disease is rare because of agreement on "negatives" alone (6). Further, the results of two or more tests may agree completely but the tests still have a very low sensitivity and/or specificity. Nevertheless, series testing is often the only practical way of estimating the specificity of a test. It should not be used, in general, to assess the sensitivity of a test.

Accuracy and Precision of a Test

An accurate test has the ability to give a true indication of the nature and quantity of the substance or objects being measured. An inaccurate test can be of value provided that the "errors" or "biases" are identifiable and consistent. For example, an inaccurate test could underestimate the true number of WBC's by 30%. Nevertheless, if the screening level were lowered by 30% from 12,948 to 9,063 cells, then the sensitivity and specificity of the test would be unaffected. However, if two tests with different accuracies or if the same test with different accuracies under different conditions were used in a testing program(s), then the sensitivity and/or specificity of the testing program(s) would differ, unless adjustment were made for these errors. On the other hand, a 100% accurate test may be of no value in a testing program because of too great an overlap in the distribution, between healthy and diseased animals of the parameter being measured. An example would be attempting to discriminate between animals with lead poisoning and healthy animals, based on the results of an accurate WBC count.

A precise test has the ability to give consistent results in repeated determinations in the same sample or animal. Precision is usually measured by the standard deviation and appears to be a more important attribute than the accuracy of a test. Inconsistent results may not alter the overall average accuracy, sensitivity or specificity but such inconsistency reduces the confidence that one places in any given determination. The usual sources or reasons for lack of repeatability relate to variation in the test itself, variation in the observer or technician or variation of the parameter under study in the animal itself. Intra-animal variation is best controlled by repeated tests on the same animal (3, 6). Variation in the test itself or in the observer-technician is best remedied by continued training

and by having clearly stated and recorded observations (6).

At present, the determination of accuracy and precision are best utilized as indicators of quality control. That is, once the accuracy and precision of a test are established, it becomes a matter of judgment as to whether or not the test is suitable for use in a testing program (18). Ongoing measurements of these attributes should be performed to ensure that any inaccuracies or inconsistencies do not exceed "acceptable limits".

Statistical Techniques for the Evaluation of Tests

A variety of statistical techniques are available to aid the evaluation of tests and the selection of a particular technique should be guided by the nature of the data and the question(s) to be answered. The chi-square statistic is very useful. However, one should be aware of the assumptions inherent in and the limitations of this statistical test (15, 16). The following methods can be used to help evaluate tests and/or to compare the results of two or more tests.

To ascertain if a test has produced "significant", in a statistical sense, separation of diseased and healthy animals, use a chi-square test on data summarized in a format similar to Table I (18). Other measures of discriminating power such as Youden's Index (18) or "validity" (14) are available but each has major disadvantages. To ascertain if the sensitivities or specificities of two tests are equal, summarize the data into two 2 x 2 tables, one for the diseased group and one for the healthy group (8). Assuming that both tests were applied to the same samples or animals, the format of the 2 x 2 table should be that used in McNemar's test for correlated proportions (6, 15, 16).

Another statistic known as "Kappa" may be used to test if the results agree to an extent significantly in excess of "chance agreement" (6). This statistic is particularly useful, irrespective of the equality/inequality of the sensitivities or specificities, because it assesses whether or not the two tests are detecting the same or different animals. For example, two tests, each with a sensitivity of 80% might give similar reactions in only 60% of the diseased animals tested. Although the sensitivity levels do not differ, the degree of agreement is less than would be expected due to chance alone. The next logical step would be to study the reasons for such a phenomenon.

"Kappa" is also a useful statistic to compare agreement when sensitivity and specificity levels are not available. Here, if Kappa is significantly large in a positive direction there is the possibility that the measurements reflect the dimension they are purported to (6). In this case, one or the other test might be selected for a testing program, since the results of both tests are providing the same information. If Kappa is negative or not significantly large, then the usefulness of the test results is severely limited (6). Neither test should be used in a testing program until the reason(s) for the disagreements have been elucidated.

To determine the precision of a test, calculate the variance of the sample of test results (18). (If a test gives only dichotomous (Yes/No) results, a positive result may be coded as 1 and a negative as 0). Often, it is useful to use a hierarchical design (16) to assess the source and magnitude of variability. This allows the localization of inconsistencies to within sample variations, between sample variation, within technician variation, between technician variation, etc. Methods for reducing the variability were mentioned previously and are discussed further by other authors (6).

In summary, the four major attributes of a test have been defined and general guidelines as to their measurements are presented. A knowledge of these attributes is essential if one is to rationally interpret the results of any given test. The exact methods necessary for the evaluation of a particular test will have to be modified, hopefully within the general framework of methods presented in this manuscript.

REFERENCES

1. **BRINLEY-MORGAN, W. J., I. DAVIDSON and C. N. HERBERT.** The use of the second international standard for anti-Brucella abortus serum in the complement fixation test. J. biol. Standardization **1**: 43-61. 1973.
2. **BRINLEY-MORGAN, W. J., and R. A. RICHARDS.** The diagnosis, control and eradication of bovine brucellosis in Great Britain. Vet. Rec. **94**: 510-517. 1974.
3. **COCHRANE, A. L. and W. W. HOLLAND.** Validation of screening procedures. Br. med. Bull. **27**: 2-8. 1971.
4. **DAVIES, G.** The rose bengal test. Vet. Rec. **88**: 447-449. 1971.
5. **FERRER, H. P.** Screening for Health: Theory and Practice. Toronto: Butterworth & Co. Ltd. 1958.
6. **FLEISS, J. L.** Statistical Methods for Rates and Proportions. Toronto: J. Wiley and Sons. 1973.

7. **GRAB, B. and J. H. PULL.** Statistical considerations in serological surveys of populations with particular reference to malaria. J. trop. Med. Hyg. **77**: 222-232. 1974.
8. **KATZ, M. A.** A probability graph describing the predictive value of a highly sensitive diagnostic test. New Engl. J. Med. **291**: 1115-1116. 1974.
9. **MacKINNON, D. J.** The complement fixation test in brucellosis. Bull. Off. int. Epizoot. **60**: 383-400. 1963.
10. **MARCHEVSKY, N.** Errors in prevalence estimates in population studies: A practical method for calculating real prevalence. Zoonosis **16**: 98-107. 1974.
11. **MAY, D.** Error rates in cervical cytological screening tests. Br. J. Cancer **29**: 106-113. 1974.
12. **NICOLLETTI, P.** Further evaluations of serologic test procedures used to diagnose brucellosis. Am. J. vet. Res. **30**: 1811-1816. 1969.
13. **NISSEN-MEYER, S.** Evaluation of screening tests in medical diagnosis. Biometrics **20**: 730-755. 1974.
14. **REIF, J. S., W. H. RHODES and D. COHEN.** Canine pulmonary disease and the urban environment 1. The validity of radiographic examination for estimating the prevalence of pulmonary disease. Arch. envir. Hlth **20**: 676-683. 1970.
15. **REMINGTON, R. D. and M. A. SCHORK.** Statistics with Application to the Biological and Health Sciences. Englewood Cliffe, New Jersey: Prentice Hall Inc. 1970.
16. **SNEDECOR, G. W. and W. G. COCHRAN.** Statistical Methods, 6th Ed. Ames, Iowa: Iowa State Univ. Press. 1971.
17. **STEWART, G. T.** Predictive value of laboratory tests. Lancet II: 1010. 1974.
18. **THORNER, R. M. and Q. R. REMEIN.** Principles and Procedures in the Evaluation of Screening for Disease. Public Health Mono. No. 67, 1961.
19. **VECCHIO, T. J.** Predictive value of a single diagnostic test in unselected populations. New Engl. J. Med. **274**: 1171-1173. 1966.
20. **YERUSHALAMY, J.** Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Publ. Hlth Rep., Wash. **62**: 1432-1449. 1947.