Preventive Veterinary Medicine, 2 (1984) 463—472
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

ESTIMATING DISEASE PREVALENCE AND THE INTERPRETATION OF SCREENING TEST RESULTS

S. Wayne Martin
University of Guelph, Ontario Veterinary College, Department of Veterinary Microbiology and Immunology, Guelph, Ontario, Canada. N1G 2W1

ABSTRACT

Martin, S.W. 1984. Estimating disease prevalence and the interpretation of screening test results. Prev. Vet. Med., 2: 463-472.

Estimating disease prevalence using a test of unknown sensitivity and specificity, as defined epidemiologically, may lead to gross errors. Thus, sensitivity and specificity are defined and methods for determining them are discussed. The effects of sensitivity, specificity and the unknown true prevalence of disease on apparent prevalence and predictive values are described.

If tests have a good sensitivity (<80%) and specificity (>96%), then the apparent prevalence, will be a reasonable estimate of the true prevalence, and the predictive value of a positive test will be at least 70% if the true prevalence of disease is at least 10%.

INTRODUCTION

Tests, of a variety of types, are used to estimate the prevalence of a disease, agent, or parasite, and/or to separate animals likely to be diseased (or infected) from those likely to be healthy (or uninfected). Yet, the underlying factors -- beyond the nature of the test and the disease process -- influencing the results of a test, and hence the interpretation of the results, appear to be poorly understood.

The desirable features of a screening test have been discussed elsewhere (Thorner and Remein, 1961). In this paper, the parameters for measuring the ability of a test to correctly detect diseased animals -- its sensitivity, and the ability to correctly detect nondiseased animals -- its specificity are described. These parameters were first defined by Yerushalamy in 1947, and elaborated by Thorner and Remein in 1961. Later, Robertson (1963) described their use for evaluating bovine tuberculosis tests and recently, detailed descriptions of screening test evaluation, in veterinary medicine, have been provided by Martin (1976) and Seiler (1979).

SENSITIVITY AND SPECIFICITY

The following discussion is pertinent to all tests used to classify an animal as a member of one of two distinct populations, such as infected or not infected. A test may give a qualitative result, which is read as positive or negative; it may

0167-5877/84/$03.00

provide a discrete quantitative result, such as the highest titre at which agglutination is complete, or it may have a continuous quantitative result, such as the amount of toxin present in a tissue sample, or the diameter of dermal response to injected antigens. Quantitative test results may be converted into dichotomous results by selecting an endpoint, or level of response, above which animals are considered positive and below which animals are considered negative. Many workers include an intermediate response category, but this makes the theory of test evaluation more difficult, and often doesn't improve the application of the test in practice.

For purposes of illustration, assume that the test under consideration is a serologic test giving an endpoint titre measureable on a discrete scale, such as, 1/20, 1/40, 1/80, 1/160, etc. Further, assume that some action must be taken depending on the above titre, and for this purpose animals with a titre of equal to, or greater than, a critical titre, say 1/80, will be designated as positive, and animals with a titre less than this will be designated as negative. Finally, assume that an animal's health status may be unequivocally categorized as diseased (exposed to an agent) or nondiseased (not exposed to the agent), by means which are biologically independent of the test being evaluated.

The sensitivity (SENS) of the test is the proportion of the universe of diseased animals that give a positive test result on any one test. Negative tests on diseased animals are called false negatives. The specificity (SPEC) of the test is the proportion of the universe of nondiseased animals that give a negative test result on any one test. Positive tests on nondiseased animals are called false positives, and the animals may be designated as NVLs; ie. no visible lesions. Sensitivity and specificity are calculated as a/(a+c) and d/(b+d) respectively, in Table 1. (Sensitivity and specificity often are multiplied by 100 to express them as a percent, but for all calculations described in this paper, they should be left as proportions). Note that these definitions differ from those in immunology and toxicology, where a sensitive test is one that detects a small amount of antigen, antibody, toxin or drug. An immunologically sensitive and specific test may not be epidemiologically sensitive and specific. There is however, a close analogy between error rates in inferential statistics and screening test evaluation. For example, the proportion of false negatives (1-SENS) and the proportion of false positives (1-SPEC) are analogous to type II and type I errors respectively (Robertson, 1963).

For any test, if the substance being measured; for example, antibodies that react with brucella abortus antigen, are present in both diseased and nondiseased individuals, there will be an inverse relationship between sensitivity and specificity. In Figure, 1, the critical titre for designating an animal as positive is set at 80 or greater. Note that this titre does not completely

distinguish between the healthy and diseased groups. If the critical titre is raised, to say 160, to increase the specificity, then sensitivity decreases. If the critical titre is lowered, to say 40, to increase the sensitivity, specificity decreases. Thus, when using the test in the field, one must weigh the biologic, economic, political and humane costs of false positive and false negative results. Usually, in the control of infectious communicable diseases, the critical titre is selected to reduce the number of false negatives, at the expense of increasing false positives. As the control program nears completion, a reassessment of this strategy may be required, due to changes in the relative numbers of diseased versus nondiseased individuals. This subject will be pursued further when discussing predictive values.

For practical purposes, it may be assumed that sensitivity and specificity are independent of the true prevalence. However, this assumption should be evaluated when possible. For example, by eliminating false positive animals in a test and slaughter program, the prevalence of cross-reacting organisms may be decreased, and hence the specificity of the test may increase as the prevalence of disease decreases. Also, it should not be assumed that sensitivity and specificity levels are invariant from herd to herd or area to area in a country. Rather, selected demographic characteristics of the animals being tested should be recorded, and their association with sensitivity and specificity determined.

Evaluating the sensitivity and specificity in this manner establishes the test's ability to separate diseased and nondiseased animals on one testing. The sensitivity and specificity do not provide direct evidence of how well the test would perform if acute and convalescent sera were tested, nor direct evidence of its sensitivity or specificity if used repeatedly on the same animals. The latter relate to the discriminatory power of a particular testing program, or regime, not just the test. Of course, the same general strategy for assessing sensitivity and specificity can be applied to program evaluation (Gray and Martin, 1980).

The sensitivity of a test at the herd level has been investigated by a number of workers. It involves not only the sensitivity and specificity of the test, but also the number of animals included in the sample (n), and the proportion of the herd that is tested. A formula for calculating the sensitivity at the herd level is:

1 - ((TP x (1-SENS) + (1-TP) x SPEC) * n)

where * n means to the nth power (Adler and Wiggins, 1973), and TP is the true prevalence of disease, expressed as a proportion.

PREDICTIVE VALUE

Suppose that the test is applied to biologic specimens from a sample -- hopefully selected randomly -- of animals in the population. The predictive value

of a positive test (PV+) is the proportion of animals with a positive test that have the disease, or other outcome of interest. The predictive value of a negative test (PV-) is the proportion of animals with a negative test that do not have the disease. These are calculated as a/(a+b) and d/(c+d) respectively, in Table I.

Predictive values are a function of (affected by) sensitivity, specificity and true prevalence of disease, and examples of this are shown in Table II. For a positive test result this relationship may be represented as

PV+ = TP X SENS / (TP X SENS + (1 - TP) X (1 - SPEC))

The predictive value of a positive test decreases directly, as the true prevalence of disease decreases (Table II).

The predictive value describes how well the test may function in any given population, but provides little information on how well the test can discriminate between diseased and nondiseased animals. Thus, the selection of a test should be based on sensitivity and specificity considerations, not predictive value (Robertson, 1973).

APPARENT PREVALENCE

The apparent prevalence (AP) is the proportion of animals tested that give a positive test result. It is calculated as (a+b)/n, from Table I. Like predictive value it is affected by the sensitivity and specificity of the test and the true prevalence of disease. Thus, whether the apparent prevalence is a good approximation of the true prevalence, ((a+c)/n), depends on the relative numbers of false negative and false positive results. Since the true prevalence is usually less than 0.5 (50%), false positives tend to over-inflate the number of diseased animals.

As shown in Table II, if the true prevalence of disease is between 5 and 40%, and the test specificity is at least 95%, then apparent prevalence will closely approximate true prevalence. However, tests with 95-96% specificity will greatly over-estimate the true prevalence when the disease is rare, say 1-2% or less.

The relationship between true and apparent prevalence (Rogan and Gladen, 1978) may be shown as:

TP = (AP - (1 - SPEC)) / (1 - ((1 - SENS) + (1 - SPEC)))

DETERMINING SENSITIVITY AND SPECIFICITY

The theoretically correct and traditionally described way of determining sensitivity is to test a representative set of diseased animals. Approximately 100 or 200 diseased animals should give sufficiently precise estimates. For specificity, a much larger representative sample of nondiseased animals, probably 2000 or so, is required for sufficiently precise estimates.

In order to obtain meaningful results, the disease status of each animal tested must be assessed, with a high degree of certainty, by tests that are biologically independent of the one being evaluated. (For example, serologic tests are not biologically independent unless they measure unrelated substances in sera). Steps must also be taken to ensure that the diseased animals in the sample are 'representative' (typical) of the universe of diseased animals. That is, the full clinical and pathologic spectrum should be present in the same proportions as in the reference population. (Ransohoff and Feinstein, 1978). The nondiseased animals also should be typical of the 'healthy' animal population to be tested. For example, if calves less than six months of age would not normally be tested, they should not be included in the nondiseased group, for purposes of estimating specificity.

In practice, it is quite difficult to assemble a group of animals containing the full spectrum of disease and demographic factors, all tested by biologically independent tests, in the numbers required for precise estimates of sensitivity and specificity. This is particularly true for many viral diseases, since culture is both expensive and usually low in sensitivity.

This has lead some to develop modified methods of estimating sensitivity and specificity, and apparently has caused others to dispair at attempting such a task (Pietz, page 171, 1977). One way of estimating sensitivity and specificity is to compare the test results to those of a bank of other tests, the latter not being biologically independent of the test being evaluated. For example, if two commonly used tests are available, a number of animals would be tested. Animals giving positive reactions to both tests would be considered diseased, and those negative to both tests would be considered nondiseased. Animals with only one positive test result would be excluded from further evaluation. The new test is applied to the assumed diseased and nondiseased animals, or their specimens, and the sensitivity and specificity determined as before. Estimates obtained in this manner provide a crude and useful bench-mark, but are likely overly optimistic, and should be prefixed with the modifier 'relative'. The reason the estimates may be biased, is that the full spectrum of diseased and nondiseased animals are not included.

A method for estimating specificity, when the disease is known to be very infrequent, is to assume that none of the test positive animals are diseased; ie. all positive test results are false positives. Specificity is then estimated using

SPEC = 1 - AP

A practical method for estimating specificity, under more general circumstances, relies on a thorough diagnostic work-up of test positive animals, with tests that are biologically independent of the one being evaluated, to establish the

predictive value of a positive test result. Then, using the predictive value, the apparent prevalence, and outside estimates of sensitivity, one may estimate both specificity and the true prevalence of disease. In particular

TP = (AP X PV+) / SENS

and

SPEC = 1 - ((1 - PV+) X AP / (1 - TP))

When the true prevalence is below 20%, specificity estimates are not affected unduly by moderate variations in sensitivity (Table III), whereas, true prevalence estimates are affected to a greater extent (Table IV).

SOME DO'S AND DON'TS

Many workers compare the results of two or more tests, and based on this, attempt to select the most sensitive or specific test. However, such a comparison describes only the extent to which the tests agree, not which test is more sensitive or specific (Martin, 1976).

If no definitive test for the true health status of animals exists, then statistical methods for calculating maximum likelihood estimates of sensitivity and specificity are available (Quade et al, 1980). While useful, it should be noted that if biologically related tests, such as the agglutination and complement fixation tests are used, these estimates may be biased (Dohoo, 1981).

Apparently noninfected animals from infected herds should not be used as a basis for determining specificity. In general, the apparent specificity based on these animals will be a gross underestimate of the true specificity. For example, initial attempts at establishing the specificity of tests used to detect bovine brucellosis were based on animals negative to culture from infected herds. (Nicoletti, 1969). Under these conditions, the apparent specificity was approximately 75% for the tube agglutination test. Yet, the state-wide apparent prevalence was less than 3% at that time and assuming a reasonable level of sensitivity 97% (100% - 3%) would represent the minimal level of specificity. Similarly, in a recent paper on the specificity of the intradermal tuberculin test, apparently tuberculosis-free herd mates of infected cattle were used to estimate specificity (Dodds, 1978). The highest apparent specificity, on this basis was 83%, suggesting that at least 17 out of every 100 tuberculosis-free animals tested would give positive reactions to the intradermal tuberculin test. A much better estimate of 96 to 97% was provided in the original paper (O'Reilly and McClancy, 1975) by noting the apparent prevalence of tuberculosis in herds with no confirmed cases. (SPEC = 1 - AP).

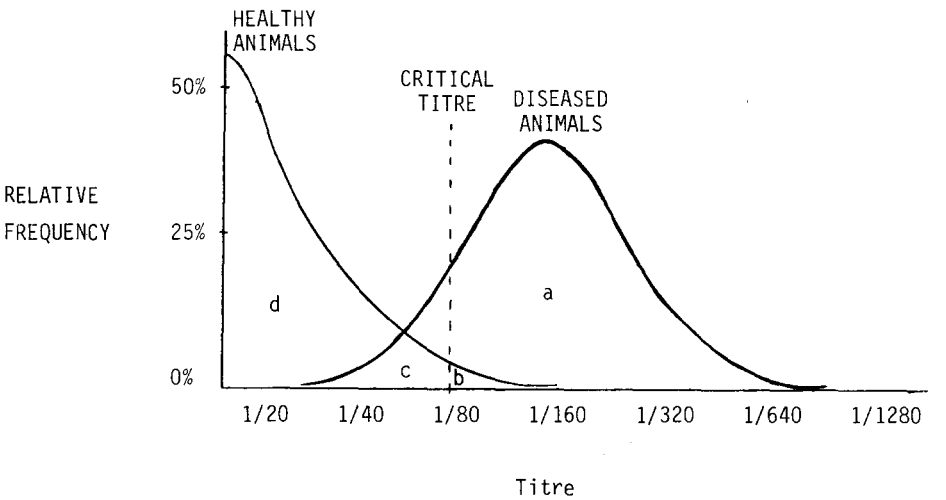
Thus, in attempting to estimate specificity based on observed values of predictive value, apparent prevalence and outside estimates of sensitivity, the predictive value and apparent prevalence statistics must relate to the entire

population under consideration, not just known infected herds. This in turn means that reactors from a number of sources must be exhaustively examined for the presence of the disease, agent or parasite. Francis et al, 1978 provide a good discussion of the practical considerations in determining sensitivity and specificity, as well as data on various bovine tuberculin tests.

Based on the previous discussion, it should be obvious that valid estimates of sensitivity and specificity are a prerequisite to the broader objective of this session; namely, estimation of disease prevalence.

FIGURE 1

The distribution of titres to agent X in a sample of healthy and diseased animals



a = True Positive
b = False positive
c = False Negative
d = True Negative

TABLE 1

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. Small case letters are used to indicate numbers in a sample

Sensitivity (SENS) = A/(A+C)
Specificity (SPEC) = D/(B+D)
True prevalence (TP) = (A+C/N)
Apparent prevalence (AP) = (A+B)/N
Predictive value (PV+) = A/(A+B)

TABLE 2

The predictive value of a test, and apparent prevalence of disease, according to the true prevalence of disease, and the sensitivity and specificity of the test.

True Prevalence	Sensitivity	Specificity	Predictive Value	Apparent Prevalence
5	80	96	51.3	7.80
	95	96	55.6	8.55
	80	99	80.8	4.95
	95	99	83.3	5.70
10	80	96	70.0	11.60
	95	96	72.5	13.10
	80	99	89.9	8.90
	95	99	91.4	10.40
20	80	96	83.3	19.2
	95	96	85.6	22.2
	80	99	95.2	16.8
	95	99	96.0	19.8
40	80	96	93.0	34.4
	95	96	94.1	40.4
	80	99	98.2	32.6
	95	99	98.5	38.6

TABLE 3

The estimated specificity of a test based on predictive value, apparent prevalence and estimated sensitivity

Apparent Prevalence	Predictive Value 40%	60%	80%
5%	96.91-96.93*	97.91-97.93	98.94-98.95
10%	93.64-93.72	95.63-95.71	97.74-97.80
15%	90.16-90.36	93.11-93.33	96.38-96.54
20%	86.45-86.83	90.34-90.77	94.81-95.14
30%	78.28-79.23	83.85-85.00	90.87-91.82
40%	68.89-70.81	75.65-78.18	85.26-87.59
50%	58.61-61.43	65.00-70.00	76.67-82.00

* (SENS = 70% - SENS = 90%)

TABLE 4

The estimated true prevalence of a disease based on predictive value, apparent prevalence and estimated sensitivity of a test

Apparent Prevalence	Predictive Value 40%	60%	80%
5%	2.22-2.86*	3.33-4.29	4.44-5.71
10%	4.44-5.71	6.67-8.57	8.89-11.43
15%	6.67-8.57	10.00-12.86	13.33-17.14
20%	8.89-11.43	13.33-17.14	17.78-22.86
30%	13.33-17.14	20.00-25.71	26.67-34.29
40%	17.78-22.86	26.67-34.29	35.56-45.71
50%	22.22-28.57	33.33-42.86	44.44-57.14

* (SENS = 70% - SENS = 90%)

REFERENCES

Adler, H.E. and Wiggins, A.D. 1973. Interpretation of serologic tests for Mycoplasma gallisepticum, World Poul. Sci. J. 29: 345

Dodd, K., 1978. Estimation of the sensitivity, specificity and predictive value of the intradermal tuberculin test. Irish vet. J. 32: 87-89.

Dohoo, I.R. 1981. Letter to the editor: Re: Effects of misclassification on statistical inferences in epidemiology. Am. J. Epidemiol. 118: 485-486.

Francis, J., Seiler, R.J., Wilkie, I.W., O'Boyle, D., Lumsden, J.J. and Frost, A.J. 1978. The sensitivity and specificity of various tuberculin tests using PPD and other tuberculins. Vet. Rec. 103: 420-435.

Gray, M.D. and Martin, S.W. 1980. An evaluation of screening programs for the detection of brucellosis in dairy herds. Can. J. comp. Med. 44: 52-60.

Martin, S.W. 1976. The evaluation of tests. Can. J. comp. Med. 41: 19-25.

Nicoletti, P. 1969. Further evaluations of serologic test procedures used to diagnose brucellosis. Am. J. vet. Res. 30: 1811-1816.

O'Reilly. L.M. and McClancy, B. 1975. A comparison of the accuracy of a human and a bovine tuberculin PPD for testing cattle with a comparative cervical test. Irish vet. J. 29: 63-70.

Pietz, D. 1977. In , Bovine Brucellosis An International Symposium. Edited by Crawford R.P. and Hidalgo R.J. Texas A & M University Press.

Quade, D., Lachenbruch, P.A., Waley, F.S., McClish, D.K. and Haley, R.W. 1980. Effects of misclassifications on statistical inferences in epidemiology. Am. J. Epidmiol. 111: 503-515.

Ransohoff, D.F. and Feinstein, A.R. 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New Eng. J. Med. 299: 926-930.

Robertson, T.G. 1963. Diagnosis of bovine tuberculosis 1. The evaluation of tuberculin tests. New Zealand Vet. J. 11: 6-10.

Rogan, W.J. and Gladen, B. 1978. Estimating prevalence from the results of a screening test. Am. J. Epidemiol. 107: 71-76.

Seiler, R.J. 1979. The non-diseased reactor: Considerations on the interpretation of screening tests. Vet. Rec. 105: 226-228.

Thorner, R.M. and Remein, Q.R. 1961. Principles and procedures in the evaluation of screening for disease. U.S. Publ. Hlth. Mono. No. 67.

Yerushalamy, J. 1947. Statistical problems in assessing methods of medical diagnosis with special reference to X-ray techniques. U.S. Publ. Hlth. Rep. 62: 1432-1449.