What do diagnostic tests really tell us ?

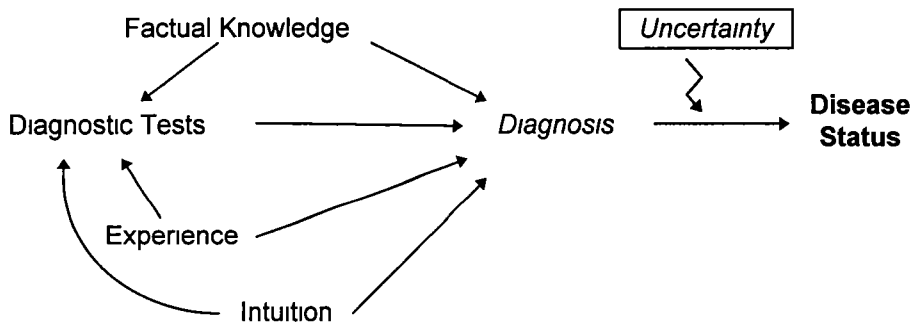
D. U. Pfeiffer

Department of Veterinary Clinical Sciences, Massey University, Palmerston North, New Zealand

Diagnosis and the veterinary profession

The duties of the veterinary profession include "to maintain and enhance the health, productivity and well-being of all animals" and "to prevent and relieve animal suffering" (from: A guide to professional conduct for veterinarians in New Zealand. New Zealand Veterinary Association, May 1991). In order to fulfill this duty the veterinarian has to be able to diagnose disease or production problems as well as identify possible causes. Diagnosis is the basis for a decision, that is to treat (or implement a program) or to do nothing, to further evaluate or to wait. The tools which the veterinarian uses to come to a diagnosis include factual knowledge, experience, intuition and diagnostic tests (see Figure 1). Correct use of these four mechanisms maximises the probability of a correct diagnosis. The uncertainty with regard to the effect of a treatment on a patient's health made the ancient greek call medicine a *stochastic art*. Clearly the main task of any doctor is to deal with the uncertainty of both diagnosis and the outcome of treatment. It has been shown in studies of the medical profession that fear of personal inadequacy and failure in reaction to this uncertainty is a common characteristic among physicians. This has become an even more important problem as our society becomes increasingly specialized and technological, relying on science rather than religion or magic to explain uncertainties (Gerrity *et al* 1992).

Figure 1 Factors in the process of diagnostic reasoning



The diagnostic process

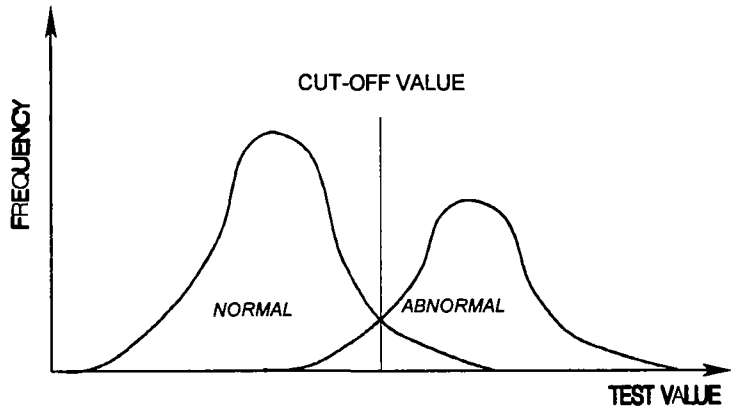
With regard to diagnosis the objective is to determine if an individual animal is in a normal condition or not. This does not necessarily mean that it has a disease. The diagnosis could also be poor reproductive performance. During the diagnostic process every practitioner assesses the probability that a particular animal has a disease or not using a number of diagnostic tools. These diagnostic tools or tests are ordinarily thought to be performed in a laboratory. But they also include methods applied during

clinical examination or x-rays. During the process this probability is constantly revised based on newly available diagnostic data. Knowledge of the prevalence of the disease in the population is usually the starting point. It provides the initial probability that the animal has the disease in question. The following clinical examination allows revision of this probability, which is then revised again after each diagnostic procedure.

The use of diagnostic tests

A diagnostic test provides a more or less objective way of reducing the uncertainty about the true condition of the animal. The outcome of a test usually will be interpreted as a dichotomous variable (yes/no type variable) indicating if the animal has the disease or not. This is straight forward for example if the test produces a result indicating bacteriologically the presence or absence of a particular organism which is already a dichotomous result. By contrast, measuring for example serum antibody levels or somatic cell counts in milk the test result will be measured on a continuous scale which in turn has to be converted into a dichotomous variable. Hence, there will be uncertainty in relation to which "cut-off" level signifies abnormality as distinct from normality (see figure 2). In reality there would be a gradual progression from healthy to diseased. With increasing test values the likelihood that the condition of the animal is abnormal would become more and more certain. The test results which range around the "cut-off" value provide most problems to the practitioner.

Figure 2. Distribution of test values for normal and abnormal populations



Evaluation of test performance

This uncertainty about the relationship between diagnostic test result and the true status of the animal can be quantified. The performance of a test can be evaluated to provide information about its *accuracy* and *bias*. The test's *accuracy* describes the closeness of the test result to the true clinical state. It is the proportion of all test results, both positive and negative, which are correct. *Accuracy* depends on disease prevalence in the population and is therefore not appropriate for test comparisons. *Bias* is any systematic deviation of the results from the true clinical state. This can result from the clinicians subjectivity where the interpretation of a particular test result is influenced by the result of a prior clinical examination. A test's *precision* (reproducibility, repeatability) identifies the degree to which a series of results based on the same sample fluctuates around a central measurement which does not have to

be accurate. In other words *precision* tells us how likely a test is to give the same result for repeated tests of the same sample.

Any such test evaluation requires the comparison of test results with a *gold standard* The *gold standard* is the true condition of the animal which is most often impossible to define with 100% accuracy. Post-mortem examination can be considered the ultimate *gold* standard, but there are disorders which cannot be confirmed through necropsy. These include subtle biochemical or neurologic alterations which can only be diagnosed in the living animal. There can be substantial disagreement between experts about what can be considered the *gold standard.* For example, in the case of mastitis, some might consider the presence of a pathogen OR the presence of an inflammatory response in the udder as the *gold standard*, whereas others would consider the presence of BOTH markers as the *gold standard.*

Given two groups which have been divided into diseased and non-diseased animals using the *gold standard,* the performance of a diagnostic test can be evaluated by applying it to each animal from the two groups. The disease spectrum found in the population these two groups have been drawn from should be as similar as possible to the disease spectrum in the population the diagnostic test is going to be used in. This can be quite difficult as many evaluations are done using animal populations kept by pharmaceutical companies under controlled conditions.

The characteristics of a diagnostic test are quantified using a two-by-two table based on the four different combinations between true disease status and diagnostic test result (see Table 1).

Table 1: Two-by two table for evaluation of diagnostic tests

	diseased	**not diseased**
test positive	a	b
test negative	c	d

Properties of diagnostic tests

Two measures, *sensitivity* and *specificity*, are traditionally described as the characteristics of a test. They provide information about how a test performs in animals with known disease status. Both measures can be used as a criterion to come to a decision on which test to use.

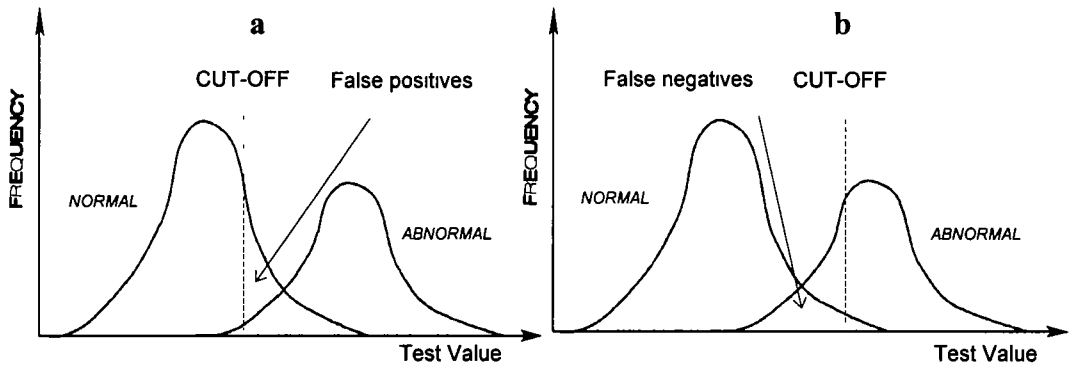
A test's *sensitivity* is defined as the proportion of animals that actually have the disorder which test positive. In other words the *sensitivity* of a diagnostic method gives us an idea how likely a test is to detect truely diseased animals. A test which is highly sensitive will rarely miss animals which do have the disease (*false negatives*). It is calculated as described in Figure 4.

Specificity on the other hand is defined as the proportion of animals which do not have the disorder which test negative. Hence, *specificity* tells us how likely a test is to correctly identify non-diseased animals as not having the disease. A test with high

specificity will rarely misclassify non-diseased animals as diseased (*false positives*). Figure 4 shows the method of calculation of *specificity*.

For diagnostic tests producing overlapping distributions of continuous data such as antibody titers or somatic cell counts for normal and abnormal animals, *sensitivity* and *specificity* can be varied by changing the cut-off value differentiating positive and negative results. An increase in *sensitivity* typically results in more *false positives* and therefore reduced *specificity* (see Figure 3a). An increase in *specificity* will produce more *false negatives* and therefore a lower *sensitivity* (see Figure 3b). Depending on the relative costs (this could be for example in financial terms or the consequences of introducing a disease such as rabies into New Zealand) of *false negatives* and *false positives*, appropriate cutpoints can be chosen to minimise one or the other.

Figure 3: Effect of changing cut-off values on sensitivity and specificity



Depending on the reasons for using a diagnostic test, knowledge about a test's properties can be useful in selecting an appropriate test. If a test is conducted with the primary objective to make sure that an animal does not have the disease (*to rule out* the target disorder), then only minimal numbers of *false negatives* are acceptable, which means that the *sensitivity* should be high. This could be the case if a farmer is trying to protect his stock from an infectious diseases through purchase of infected animals. In this case *false negative* test results could be disastrous for the farmer, whereas a *false positive* would not be of major concern as he/she would not buy the stock. If the objective is to make sure that a test positive animal really has the disease (*to rule in* the disease), then only few *false positives* will be acceptable and therefore *specificity* should be high. This is the perspective of a farmer who is trying to sell stock or become accredited as being free from a certain disease. Here *false positives* clearly would be extremely undesirable.

As *sensitivity* and *specificity* depend to a large extent on the disease spectrum found in the population used for the test evaluation, the estimates for a particular test reported in the literature can vary significantly. The disease spectrum can be quite different if animals used for test evaluation are exposed to natural or experimental infection. A number of other factors including animal characteristics such as breed, age and sex or environmental conditions can modify the immune response and/or pathogenesis and therefore result in different disease spectra found in populations.

Sensitivity/specificity estimates based on particular population are also subject to chance variation which depends on the size of the sample of animals which has been used to evaluate the test. The degree of precision of the estimates can be defined using *95% confidence intervals* (see Fletcher *et al* 1988). A narrow confidence interval indicates that the observed *sensitivity/specificity* is close to the true *sensitivity/specificity*.

Interpretation of test results

Once the results of a diagnostic test are available, *sensitivity* and *specificity* are no longer of primary importance, because both measures require knowledge about the true condition of the animal. Instead, the true condition of the animal is typically what the clinician is trying to find out using a diagnostic test. What the clinician requires is an estimate of the probability that the particular test result represents the true condition of the animal. This can be done using *predictive values* or *post-test* (*posterior*) probabilities.

The *predictive value of a positive test* is the proportion of test positive animals which truly have the disease. In other words it tells us what the chances are that given a positive test result, a particular animal truely has the disease in question. The *predictive value of a negative test* represents the proportion of test negative animals which do not have the disease. This gives us the probability that a negative test result for a particular animal really means that this animal does not have the disease.

Both *predictive values* require knowledge of the *sensitivity* and *specificity* of the diagnostic test and of the *true prevalence* of the condition in the population. The more sensitive a test is the better will be its *negative predictive value* (the more confidence a clinician can have in a negative test result). A more specific test will have a better *positive predictive* value. Predictive *values* will vary with disease *prevalence*. This means that the interpretation of positive as well as negative test results should be varied between populations, according to the *prevalence* found in the particular population. For example if disease *prevalence* is low such as during the final phase of a disease control campaign, even with a very specific test *positive predictive values* will decrease and many animals will be culled which do not have the disease. In the example presented in Table 2 given a 6% *prevalence* the *predictive value of a positive test* result is 0.50, which means that given a positive test result there is a 50% chance that this animal has been correctly classified as diseased. Similarly, in a high *prevalence* situation even a very sensitive test cannot prevent that there will be *false positives* resulting in a poor *negative predictive value*. Table 3 presents a scenario using the same test characteristics as the test used in Table 2, but this time *prevalence* in the population is assumed to be 60%. The *predictive value* of a *positive* test has gone up to 0.96, whereas the *predictive value* of a *negative* test result has gone down to 0.86. Diagnostic tests should therefore not be compared on the basis of their *predictive values*. They only represent the probability that an individual animal is correctly classified given a certain test result.

Table 2. Effect of low prevalence on predictive values

95% specificity 90% sensitivity **6% prevalence**

	True disease status			
Test result	**Disease**	**No disease**	**Total**	**Predictive value**
Positive	5	5	10	0 50
Negative	1	89	90	0 98

Table 3: Effect of high prevalence on predictive values

95% specificity 90% sensitivity **60% prevalence**

	True disease status			
Test result	**Disease**	**No disease**	**Total**	**Predictive value**
Positive	54	2	56	0 96
Negative	6	38	44	0 86

For estimation of *predictive* values, a correct estimate of *prevalence* is at least as or may be even more important than *sensitivity/specificity*, because the latter typically do not vary over such a wide range as *prevalence* does. To achieve a high *positive predictive value* it is desirable to work with a population or subpopulation which has a high *prevalence*. *Prevalence* is also called *prior* or *pretest probability*. There are a number of ways how *prevalence* or *prior probability* can be increased before using a diagnostic test. For example a clinician would only apply the test to a group of animals where he/she suspects that they might have the disease based on clinical information. Hence, appropriate interpretation of history and physical examination can be used to increase the likelihood of disease in the animal before it is tested compared with an animal randomly selected from the population. In this situation the test is used as a *diagnostic test* which is different from its use as a *screening test*, where it is applied to animals from a population without prior clinical investigation (e.g. during disease control scheme).

If a test is used as a *screening test* to estimate *prevalence* of a condition in a population there will be false positive and false negative animals and not necessarily in equal numbers. Hence, only the *apparent prevalence* is estimated which depending on a test's sensitivity and specificity does not have to equal the *true prevalence*. Taking the example in Table 2 *apparent prevalence* based on the results of the test would be 10%, whereas *true prevalence* was 6%.

Figure 4: Calculation of test characteristics

	DISEASE	NO DISEASE	TOTAL
TEST POSITIVE	a	b	a + b
TEST NEGATIVE	c	d	c + d
TOTAL	a + c	b + d	N

$$\text{SENSITIVITY} = \frac{a}{a+c}$$

$$\text{SPECIFICITY} = \frac{d}{b+d}$$

$$\text{POSITIVE PREDICTIVE VALUE} = \frac{a}{a+b}$$

$$\text{NEGATIVE PREDICTIVE VALUE} = \frac{d}{c+d}$$

$$\text{APPARENT PREVALENCE} = \frac{a+b}{N}$$

$$\text{TRUE PREVALENCE} = \frac{a+c}{N}$$

$$\text{ACCURACY} = \frac{a+d}{N}$$

Glossary

Accuracy — proportion of correct test results (see Figure 4)

Apparent prevalence — estimate of proportion of diseased animals in population based on test result (see Figure 4)

Bias — systematic deviation from true clinical measurement

Diagnostic test — used by clinician during a general diagnostic work-up to come to a diagnosis about the condition of an individual animal

Gold standard — method for assessing the true disease status in a test evaluation

Post-test (posterior) probability — probability that an individual animal has the disease after a diagnostic test has been applied

Pre-test (prior) probability — probability that an individual animal in a sample from a population has the disease before a diagnostic test has been applied (same as *prevalence*)

Precision — degree of fluctuation of repeated tests based on the same sample around central measurement

Predictive value of a *negative* test — likelihood that an animal with a *negative* test result truely does not have the disease (see Figure 4)

Predictive value of a *positive* test — likelihood that an animal with a *positive* test result truely has the disease (see Figure 4)

Screening test	test which is applied to animals from a population without prior clinical investigation
Sensitivity	proportion of *diseased* animals which test is able to correctly classify as diseased (see Figure 4)
Specificity	proportion of *non-diseased* animals which test is able to correctly classify as non-diseased (see Figure 4)
True prevalence	true proportion of diseased animals in population based on *gold standard*

References

Gerrity,M.S., Earp,J.L., DeVillis,R.F. and Light,D.W. 1992: Uncertainty and professional work: Perceptions of physicians in clinical practice. *American Journal of Sociology*, 1022-1051.

Fletcher,R.H., Fletcher,S.W. and Wagner,E.H. 1988: *Clinical epidemiology - the essentials*. 2nd Ed., Williams & Wilkins, Baltimore, Maryland, U.S.A., 246pp.

Further Reading

Dawson-Saunders,B. and Trapp,R.G. 1990: *Basic and clinical biostatistics*. Prentice-Hall Int., London, 329pp.

Fletcher,R.H., Fletcher,S.W. and Wagner,E.H. 1988: *Clinical epidemiology - the essentials*. 2nd Ed., Williams & Wilkins, Baltimore, Maryland, U.S.A., 246pp.

Kennedy,D. (technical editor) 1990: *Epidemiological skills in animal health.* Proceedings 143, Post Graduate Committee in Veterinary Science, University of Sydney, Sydney, Australia, 409pp.

Kraemer,H.C. 1992: *Evaluating medical tests*. Sage publications, Newbury Park, California, 294pp.

Martin,S.W., Meek,A.H. and Willeberg,P. 1987: *Veterinary Epidemiology - Principles and Methods*. Iowa State University Press, Ames, Iowa, U.S.A., 343pp.

Sackett,D.L., Haynes,R.B., Guyatt,G.H. and Tugwell,P. 1991. *Clinical epidemiology - A basic science for clinical medicine*. 2nd Ed., Little, Brown and Company, Boston, U.S.A., 441pp.

Smith,R.D. 1991: *Veterinary clinical epidemiology*. Butterworth, Boston, U.S.A., 234pp.