How Should We Use Diagnostic Tests in Practice?

R S Morris and R Jackson
Department of Veterinary Clinical Sciences,
Massey University, Palmerston North

What Do We Want from Tests?

We want a test which will give the correct answer in every case (100% sensitivity and 100% specificity) throughout the course of a disease or condition, and will give high repeatability if used again on the same animal after a short time interval If possible, the test should achieve these goals on individual animals, although a test which achieved these aims on a herd basis may in some cases be satisfactory (ie it would classify herds correctly, although not necessarily each animal within the herd)

There is no such test! ***All*** of the tests we use fall short of this goal, and most of them fall far short However veterinarians commonly interpret all laboratory test results as if tests performed perfectly In part this is due to misplaced trust, but it is also due to not being fully aware of the specific ways in which test results should be interpreted This paper will explain how better interpretation of test results can be made

It will answer questions such as

How do we judge the value of a particular test?
Can we improve the accuracy of tests by combining them?
If we already know other information about the likelihood of a disease occurring in an animal, should we allow this to influence test interpretation?
How do we explain to clients when a test proves to have given false results?
What approach should we use in trying to get the best value out of tests for individual animals, and for herd disease control?

Perfect and Imperfect Tests

All methods of disease diagnosis are imperfect in telling us the "true state of nature" with regard to a particular animal or herd - some are just a lot more imperfect than others We have to have some benchmark against which to make judgments about the value of new tests, so for each disease or condition we define one particular method of diagnosis to be "the gold standard" against which we measure other tests Usually it is some method of diagnosis which is more time-consuming, difficult or expensive to conduct than the test we use in practice (if it wasn't, it should have been the test of choice for field use!) As technology improves, the gold standard for a particular case may change, although we should be very cautious about assuming that just because a diagnostic technique is more complicated and technically advanced, it is necessarily better In tuberculosis diagnosis, the gold standard is usually autopsy examination But which should it be - gross lesions found in an abattoir examination, gross lesions found in a detailed autopsy, gross plus histopathological examination of apparently lesion-free animals, culture of *M bovis* from an examination of lesioned animals, culture of *M bovis* from an examination of all animals in the test group, or positive results on PCR (polymerase chain reaction test) from all animals in the group? Moving through the list, each of these techniques is likely to give us a higher and higher proportion of animals classified as positive - which is closest to the truth? Is it always necessarily the one which gives the largest number of positives?

Ultimately, the people who are experts in a particular disease have to decide what is an appropriate gold standard, and then each of the other tests can be measured against that Although they don't always agree, in most cases some degree of consensus emerges over time, and then we can get sensitivity and specificity results on a test by comparing it with the gold standard If there is either no gold standard for a particular disease or we cannot afford to evaluate a new test against the accepted gold standard, we can still make a (considerably less precise) assessment of the value of a test by comparing it with a second test generally

considered to be highly sensitive and specific However the validity of this depends on how close the second test approaches gold standard status, and in comparing the two tests we must take into account the fact that they will agree in a substantial proportion of cases purely by chance Hence the methods we use to assess tests takes this into account So for every significant test technique we use, from tuberculin testing to serology to clinical chemistry, we need to think in terms of the sensitivity and specificity of the procedure in diagnosing the disease of interest, and ask for that information on each test It is not enough to be assured that test X is a "good test" or a "reliable test" - good for what?

If you are a farmer wanting at any cost to save a valuable animal from being declared positive for a particular disease, then you want close to perfect specificity, even if it means poor sensitivity, which will mean inevitably leaving infected animals in some herds If you are a quarantine veterinarian wanting to be as confident as possible that a group of animals is free of a highly infectious disease before allowing them to enter New Zealand, you will go for a highly sensitive test regardless of the fact that it will exclude some uninfected animals from entering the country

Between these two extremes is the territory of the epidemiologist - finding the right testing strategy for the particular disease control objective which balances the competing considerations - because until you decide on your objective we can't tell you what test to use In this paper we will outline the process by which test selection and application can logically be approached for different objectives

Choosing the Cut-Off Point

Relatively few tests have a clearcut yes/no answer One example is bacteriological culture, where either the organism grows on the plate or it doesn't But a negative culture certainly doesn't necessarily mean that the animal was not infected - perhaps if we had used a selective culture medium or prepared the sample differently we would have got a different result So even in this case we have a decision to make about how sensitive or specific we want the testing procedure to be

But for most tests used by practitioners the result is measured on a scale of some kind, and a decision has to be made about where the dividing line will be between positive and negative results The common view is that the separation is clearcut The reality is that test results for diseased and normal animals overlap for every test and whatever cut-off point is chosen, some animals will be wrongly classified As the cut-off point is shifted, the only choice is between reducing false positives by increasing the number of false negatives, or vice versa. Figure 1 shows what everyone would like to be the case, while Figure 2 shows the real situation

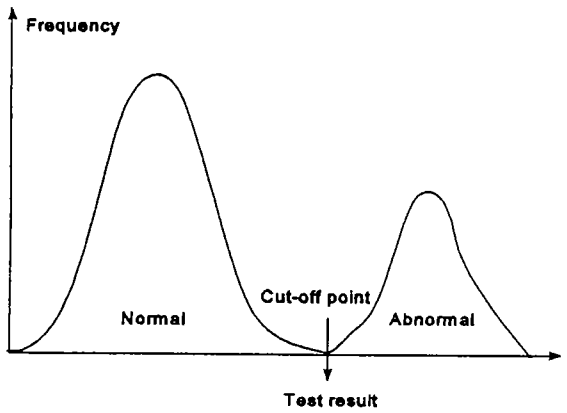

Figure 1. The common view of diagnostic tests

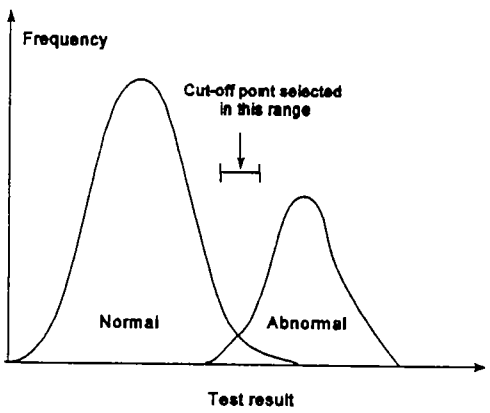

Figure 2. The true situation with diagnostic tests

Although laboratory staff aim to select a cut-off point which is likely to correctly classify as high a proportion of animals as possible, it is impossible for them to get it completely right, and anyway the set of sera you

supplied to them may have important differences in test response from the population on which they originally standardised the test So the choice of cut-off point is a critically important decision, which will be influenced both by how the test performed in its original evaluation, plus the views of the laboratory person on the relative importance of false positives and negatives for that test in those particular circumstances It is not fixed, and so for many tests the cut-off point may vary between laboratories in different parts of the world - hence caution is required in interpreting published information reported as positive/negative rather than primary results Even primary results are influenced by the calibration of the test procedure in many cases, so will vary between laboratories

The test result is also influenced by the characteristics and history of the animal population being tested, and sensitivity and specificity of a test may vary between populations for a whole variety of reasons, even if they have the same prevalence of the disease (which is an important factor in determining the predictive value of tests, as discussed below) One example of variation in test performance between populations is the occurrence of non-specific reactions to the tuberculin test due to environmental mycobacteria, which substantially reduces the specificity of the test in areas where these bacteria are common

Combining Tests - Does It Give a Better Answer?

The common response to receiving a test result which does not fit expectations is to do an alternative test on the grounds that the combination must be more reliable than just a single test Hence screening tests have become common in investigation of some diseases, and in clinical chemistry in particular the "battery of tests" is assumed to be more informative than testing for just one metabolic factor Unfortunately, the truth of test interpretation is not so comfortable, and there is no escape from the consequences (both positive and negative) of "doing another test to sort it out"

Where two or more tests are used, they have to be interpreted either in parallel or in series. The order in which the tests are done does not influence this - it is simply an approach to final interpretation Parallel interpretation means that an animal is considered to have the disease of interest if any one or more of the tests produce a positive result This increases the sensitivity of the total test procedure, but reduces its specificity Series interpretation of tests means that an animal is considered to have the disease of interest only if all the tests used produce a positive result This increases the specificity of the test procedure, but reduces its sensitivity

Thus series interpretation reduces the number of false positives but increases the number of false negatives Parallel interpretation does the reverse. This is simply a consequence of the way in which the tests are interpreted, and is inescapable, although some authors make out that they combine tests in a way which increases both sensitivity and specificity Commonly such claims relate to series testing, where one test is used as a screening test and a second as the definitive test Claims of superior test performance are sometimes made for the definitive test alone when it has been interpreted in series with the screening test, and hence only positives to the first test are subjected to the second test. This is quite unacceptable, because the population which goes on to the second test is very biased - sensitivity and specificity can only be considered in relation to the total tested group, and in this case the laws of test performance will apply as usual If more than one test is done, each should also be read "blind" (ie without knowing the result of the other test), since it is difficult to avoid being influenced by another result if it is known, despite good intentions

The following mythical example for mastitis control illustrates the effect of different approaches to test interpretation

Test Result		True Situation	
High cell count	*Pathogens isolated*	*Truly infected*	*Truly negative*
+	+	60	10
+	-	20	20
-	+	10	40
-	-	10	930
Total		100	1000

Without taking space to show the calculations, for high cell count interpreted alone, sensitivity is 80% and specificity 97%. For isolation of pathogens alone, sensitivity is 70% and specificity 95% If the tests are interpreted in parallel, the calculated sensitivity rises to 90% but the specificity falls to 93% If the tests are interpreted in series, the calculated sensitivity falls to 60% but the specificity rises to 99% Thus combining tests can improve the value of the final conclusion if one test has performance characteristics which compensate for limitations of the other, such as using a cheap and sensitive screening test followed by a much more expensive and highly specific definitive test on samples positive to the screening test, and interpreting the results in series However in doing so it is important to realise that ***you cannot do this without sacrificing either sensitivity or specificity, depending on which method of interpretation you choose!*** Be very wary of any combined test procedure which claims not to suffer from this problem - it is likely that some of the results are being reported only for a selected sub-group of the total population, which invalidates the conclusion Note that because specificity is usually higher than sensitivity, series interpretation produces a large drop in sensitivity relative to the gain in specificity The change is not so marked for parallel interpretation - but series interpretation is far more commonly used! Although strictly the sensitivity and specificity of the combined tests should be determined directly rather than by multiplication of the individual test values, the calculated figures give a reasonable approximation to the figure expected in a field evaluation of the procedure

The same principles apply to re-testing of animals with the same test on two occasions This is usually done to resolve the state of animals which test positive in a herd considered uninfected with the disease Test interpretation is therefore in series, and specificity will be increased but sensitivity reduced, hence the animal is more likely to be declared negative at the re-test, regardless of its true infection state This is satisfactory if the animal is almost certain to be a false positive, and it is unlikely that this animal might be the first case of the disease in the herd. However if we re-test "singleton reactors" routinely in an area where a disease is spreading to new herds, we guarantee that we will miss a substantial proportion of newly infected herds until infection has spread more widely So be cautious about the circumstances under which you use this approach

One special case of re-testing which is best called sequential testing, is where the purpose is to detect a rise in titre due to recent infection - sometimes called acute/convalescent sampling in human medicine This is *not* series testing, since interpretation is based on the change in titre between the two samplings, not on some combination of results The first sample must be collected as early as possible in the course of the disease, and the second usually 4 to 6 weeks later A rise in titre beyond the normal biological variation (say a four-fold rise) is strong evidence that the animal became infected and seroconverted for the disease at the time of the initial examination. For investigating infectious diseases, seroconversion is a much more powerful piece of epidemiological information than a high titre at a single test This is especially true for diseases where an animal may suffer the same disease more than once or only some infected animals may manifest disease, and hence a positive titre alone may not be conclusive - such as leptospirosis.

Another problem with multiple tests is the widespread use of the "battery of tests" to identify haematological or metabolic factors which are outside the normal range and therefore suggestive of disease This is a hazard of modern laboratory equipment, where it is only marginally more expensive to run 12 tests than one specific test The technique is satisfactory if each test is evaluated and interpreted to decide whether it supports or refutes the diagnosis provisionally made on clinical grounds This may be extended to a form of parallel interpretation of test results where a group of tests which all represent possible diagnostic indicators for a particular disease are considered together to decide whether the animal has the condition We call this "pattern diagnosis", and it is useful in herd performance evaluation as well as specific disease diagnosis However it is a form of multiple test interpretation, and obeys the normal rules outlined above for multiple tests

Where the battery of tests becomes very questionable is where it is used for a fishing expedition, without some prior provisional diagnosis being made "I don't know what's wrong with it, so I'll see what the lab tests show" For any single test the normal range is usually defined so that it includes 95% of the population of normal animals - to the extent that this is known. Therefore for any single test at least 5% of normal animals will lie outside the so-called normal range If the laboratory runs 12 tests, then the probability of a normal animal having values for all 12 tests within the normal range is $(0\ 95)^{12}$, which is 54% So 46% of totally normal animals subjected to this battery of tests would produce values for one or more of the tests which were outside the normal range, and hence may be wrongly diagnosed as suffering from a disease Laboratory tests can only be used to support or refute a diagnosis, not to make the diagnosis

Combining Prior Information With Test Results

Instinctively we all interpret test results in the light of our knowledge of how common a disease is, and how extreme the test results are This is why new graduates frequently diagnose rare diseases This is a technically correct approach, and can be done more precisely through a technique called "likelihood ratio estimation" This uses an estimate of how common the disease is in the population (prevalence) in combination with the sensitivity and specificity of the test to estimate the probability that an animal with a particular test result actually has the disease.

Tests Valid for a Herd But Not an Individual Animal

Some tests have sensitivity or specificity which is so low that the test is quite unreliable on an individual animal However if the only issue of concern is whether or not the herd has the disease, then tests on a sample of animals from the herd may provide a reliable answer As usual, if interpretation is parallel, "one positive, herd positive", then a poorly sensitive test may be adequate provided that it has high specificity Series interpretation of a group of animals can overcome problems with a non-specific but very sensitive test If both are poor, look for a new test!

Prognostic Tests

Some tests may be used principally to estimate the likelihood of a particular event happening to an animal over some period of time into the future Such testing has been far more common in human medicine, for purposes such as predicting life expectancy for people with various diseases, or who have known or hypothesized risk factors for diseases, such as elevated cholesterol Such information has to be handled differently since it is not yes/no but "how long?" We use a technique called survival analysis to investigate such issues, including the predictive value of test results However this is a large and somewhat under-used area of veterinary science, which has less immediate application in practice than diagnostic testing, although it is likely to increase in the future

Diagnostic Testing for Clinical Case Management

Many of the biochemical and other tests used in companion animal diagnosis are not highly specific, and

hence if used on random animals will produce substantial numbers of false positives The predictive value of a test (either positive or negative) is highest at intermediate prevalence, not at very high or low prevalence in the tested population We rarely have problems with very high prevalence, but we do with low prevalence The predictive value of a test can be increased by limiting its use to animals considered on other grounds to be at high risk, thus raising the prevalence to a level at which the test has optimum predictive value For example, the test may be used only in young horses, or in obese dogs, or animals already showing clinical signs suggestive of the condition The test is likely to be more discriminating under these conditions. However beware of the fact that many such tests reported in journals as having high predictive value have only been tested on selected populations, such as animals referred to a University clinic. The predictive values of tests determined on such a population are likely to be much higher than in an unselected population, and hence may represent the best test performance you can expect, not the typical performance

If the test still does not meet requirements, then combining tests judiciously to improve their predictive capacity as described above offers a second approach to adopt

Diagnostic Testing for Disease Control and Eradication

Usually a disease which is the subject of a control program is at an initial prevalence in a population of less than about 20% of animals affected If the prevalence is less than about 20 to 25%, with typical values for sensitivity and specificity found for diagnostic tests for infectious diseases, the apparent prevalence (ie that estimated by the usual diagnostic test) will almost always be higher than the true prevalence (ie that estimated by a gold standard test) The lower the prevalence, the larger the gap between apparent and true prevalence Thus we will overestimate the proportion of animals infected even before we begin to take control measures, and as we progressively reduce the number of infected animals this "predictive value of a positive test" will become steadily lower Early in a disease eradication program "false positive" animals will be present but will represent only a small proportion of total positives, so if test and slaughter is being used for control, only a small proportion of animals will be wrongly identified as infected However as the control program makes progress the prevalence will fall and the proportion of positives which are false will inescapably rise and the discrepancy between apparent and true prevalence will get steadily wider Farmers have great difficulty with this idea, and pressure mounts for "something to be done" If this problem is not taken into account in advance, the disease control program can fall into disrepute just when it is making its biggest gains, and is most susceptible to a loss of momentum Yet although the problem can be marginally reduced by a change in test, it will not go away, and a lot of money can be spent on developing new tests which make little difference to the problem

The problem tends to be particularly important for tuberculosis, where farmers receive reports back giving the slaughter findings for reactor animals Unfortunately the sensitivity of abattoir inspection for tuberculosis is quite low under real-life conditions (much poorer than tuberculin testing), so a substantial number of true infected animals are classified as "no visible lesion" when they did in fact have detectable lesions of TB When this is added to the genuine false positive results, farmers mistakenly conclude that tuberculin testing is inaccurate Even worse, they erroneously draw the conclusion that unless lesions are found at slaughter they do not have TB on the farm In fact, a significant proportion of such farms do have TB, but with frequent testing (at least annual) lesions are too early to be detected without careful examination There are also cases (fortunately uncommon) in which due to infrequent testing, infected animals in a few herds become anergic before they are detected as infected, and hence spread disease within and sometimes between herds As a result of these various problems, attention in TB control tends to focus far too much on the test and its perceived deficiencies, instead of on how to eradicate infection

In the early stages of a control program test sensitivity is of paramount importance, since we are trying to reduce the number of infected animals in the population In addition, testing will be more sensitive in herds with more disease, helping us to ensure early rapid progress However as prevalence falls, specificity becomes the dominant requirement, and it may at some point be desirable to add a second test carried out in series with the first test, to increase specificity The number of animals tested per herd also has a large influence on the

accuracy of identification of infected herds - the more animals tested, the higher the sensitivity and the lower the specificity Therefore in some cases it is better to test the same number of animals regardless of herd size, to avoid the problem of continuing false positive animals in large herds The number of reactors which must be found before the herd is declared infected is also important in these circumstances - as this minimum number rises the sensitivity for detecting infected herds falls and specificity rises

The Epidemiological Approach to Test Interpretation

From the explanations given, it should be clear that there is no perfect test for diagnosing disease! The answer lies in using tests not as final answers but rather as tools which are combined with other information and skills, to work out a specific control strategy for each herd Through careful history-taking, herd examination, analysis of patterns of test positives, and appropriately timed sequential testing, a properly structured investigation of the infection status of a herd can be carried out Then by using testing strategies designed to achieve accurate herd diagnosis followed by progressive resolution of the problem in herds diagnosed as infected, successful control or eradication can be achieved In this way the limitations caused by imperfect test performance can be minimised, and client satisfaction can in most cases be maximised However because tests are imperfect, there will be occasions when a wrong conclusion will be drawn, and we must have back-up strategies to sooner or later find these cases

Conclusion

Diagnostic tests for both infectious and non-infectious diseases are an essential tool for veterinarians However in order to use them to maximum effect we need to understand their limitations as well as their virtues, and to adopt strategies which maximise the accuracy of our diagnoses and the effectiveness of our control efforts, not simply to trust tests implicitly