

# 5

---

## SCREENING AND DIAGNOSTIC TESTS

### OBJECTIVES

After reading this chapter, you should be able to:

1. Define accuracy and precision as they relate to test characteristics.
2. Interpret measures of precision for quantitative test results; calculate and interpret kappa for categorical test results.
3. Define epidemiologic sensitivity and specificity, and calculate their estimates and their standard errors (or confidence intervals).
4. Define predictive values and explain the factors that influence them.
5. Know how to choose appropriate cutpoints for declaring a test result positive (this includes using receiver operating characteristics curves and likelihood ratios).
6. Know how to use multiple tests and interpret results in series or parallel.
7. Understand the impact of using multiple tests that are not conditionally independent.
8. Describe multiple approaches to evaluating (*ie* estimating sensitivity and specificity) diagnostic tests.
9. Understand latent class models for estimating sensitivity and specificity when no gold standard exists.
10. Understand how population characteristics might affect estimates of sensitivity and specificity and be able to use logistic regression to evaluate these effects.
11. Describe the main features influencing herd-level sensitivity and specificity based on testing individual animals.
12. Describe the main features affecting the use of pooled specimens.

## 5.1 INTRODUCTION

Most of us think of tests as specific laboratory test procedures (*eg* a liver enzyme, serum creatinine, or blood urea nitrogen test). A test, more generally, is any device or process designed to detect, or quantify, a sign, substance, tissue change, or body response in an animal. Tests can also be applied at the herd, or other level of aggregation. Thus, for our purposes, in addition to the above examples of tests, we can consider clinical signs (*eg* looking for a jugular pulse), questions posed in the history-taking of a case work-up (*eg* how long since previous calving), questions in a questionnaire (*eg* about management practices) or findings at post-mortem examination of carcasses as tests. Indeed, tests are used in virtually all problem-solving activities and therefore, the understanding of the principles of test evaluation and interpretation are basic to many of our activities. Several discussions of the application and interpretation of tests are available (Greiner & Gardner, 2000a; Greiner & Gardner, 2000b).

If tests are being considered for use in a decision-making context (clinic or field disease detection), the selection of an appropriate test should be based on the test result altering your assessment of the probability that a disease does or does not exist and that guides what you will do next (further tests, surgery, treat with a specific antimicrobial, quarantine the herd *etc*) (Connell & Koepsell, 1985). In the research context, understanding the characteristics of tests is essential to knowing how they affect the quality of data gathered for research purposes. The evaluation of tests might be the stated goal of a research project or, this assessment might be an important precursor to a larger research programme.

### 5.1.1 Screening vs diagnostic tests

A test can be applied at various stages in the disease process. Generally, in clinical medicine, we assume that the earlier the intervention, the better the recovery or prognosis. Tests can be used as **screening tests** in healthy animals (*ie* to detect seroprevalence of diseases, disease agents or subclinical disease that might be impairing production). Usually the animals or herds that test positive will be given a further in-depth diagnostic work-up, but in other cases, such as in national disease-control programmes, the initial test result is taken as the state of nature. For screening to be effective, early detection of disease must offer benefits to the individual, or 'programme', relative to letting the disease run its course and being detected when it becomes clinical. **Diagnostic tests** are used to confirm or classify disease, guide treatment or aid in the prognosis of clinical disease. In this setting, all animals are 'abnormal' and the challenge is to identify the specific disease the animal in question has. Despite their different uses, the principles of evaluation and interpretation are the same for both screening and diagnostic tests.

## 5.2 ATTRIBUTES OF THE TEST *PER SE*

Throughout most of this chapter, the focus will be on determining how well tests are able to correctly determine whether individuals (or groups of individuals) are diseased or not. However, before starting the discussion of the relationship between test results and disease status, we will address some issues related to the ability of a test to accurately reflect the amount of the substance (*eg* liver enzyme or serum antibody level) being measured and how consistent the results of the test are if the test is repeated.

The terminology used in the literature to describe the evaluation of tests is not entirely

consistent (de Vet *et al*, 2006; Streiner & Norman, 2006). However, concepts that relate to the test *per se* include analytic sensitivity and specificity, accuracy and precision. Our usage of the term precision is as a general term to reflect the variability among test results.

### 5.2.1 Analytic sensitivity and specificity

The analytic sensitivity of an assay for detecting a certain chemical compound refers to the lowest concentration the test can detect. In a laboratory setting, specificity refers to the capacity of a test to react to only one chemical compound (*eg* a commonly used test in the dairy industry to identify the presence of antibiotic ( $\beta$ -lactam) inhibitors in milk). The analytic sensitivity of the test is 3 ppb for penicillin, meaning that the test can detect levels of penicillin in milk as low as 3 ppb. The test reacts primarily to  $\beta$ -lactam antibiotics but will also react with other families at higher concentrations, such as tetracyclines. Thus, the test is not specific to just  $\beta$ -lactam antibiotics. Diagnostic (epidemiologic) sensitivity and specificity depend (in part) on analytic sensitivity and specificity, but are distinctly different concepts (Saah & Hoover, 1997) and are discussed in Section 5.3.

### 5.2.2 Accuracy and precision

The laboratory accuracy of a test relates to its ability to give a true measure of the substance of interest (*eg* blood glucose, serum antibody level). To be accurate, a test need not always be close to the true value, but if repeat tests are run, the average of the results should be close to the true value.

The precision of a test relates to how consistent the results from the test are. If a test always gives the same value for a sample (regardless of whether or not it is the correct value), it is said to be precise. Fig. 5.1 shows the various combinations of accuracy and precision.

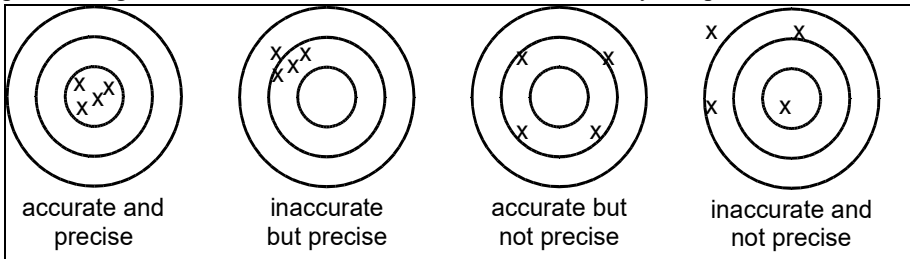


Fig. 5.1 Laboratory accuracy and precision

Results from tests that are inaccurate can only be 'corrected' if a measure of the inaccuracy is available and used to adjust the test results. Imprecision can be dealt with by performing repeated tests and averaging the results. Correct calibration of equipment and adherence to standard operating procedures are essential to good accuracy and precision; however, the details are beyond the scope of this book.

### 5.2.3 Measuring accuracy

Assessing accuracy involves running the test on samples with a known quantity of the substance present. These can be field samples for which the quantity of the substance has been determined by a generally accepted reference procedure. For example, the accuracy of an

infrared method for determining milk urea nitrogen (MUN) level in milk samples has been evaluated by comparing those results with those obtained from a ‘wet-chemistry’ analysis (Arunvipas *et al*, 2003). Alternatively, the accuracy of a test can be determined by testing samples to which a known quantity of a substance has been added. The possibility of background levels in the original sample and concern about the representativeness of these ‘spiked’ samples make this approach less desirable for evaluating tests designed for routine field use. A much more detailed description of procedures for evaluating laboratory-based tests can be found in Jacobson (1998).

#### 5.2.4 Precision and agreement

As indicated above, the term precision is used to denote variability among test results. This variability may arise in several ways. Variability among test results obtained from repeated testing of the same sample within the same laboratory is referred to as **repeatability**. Variability among test results obtained from testing the same sample in different laboratories is called **reproducibility** (World Organisation for Animal Health (OIE), 2004) and is, in part, a reflection of how easy it is to set up the test in different settings. A related concept is that of **reliability**, which refers to the ability of a test to distinguish between individuals and is not, strictly speaking, a measure of precision (see Section 5.2.6).

**Agreement** refers to how well 2 tests agree. It might refer to the level of agreement between 2 different tests for the same substance or between responses of 2 different raters who are estimating a value (*eg* 2 individuals determining the heart rate of an animal). General frameworks for evaluating agreement have recently been published (Barnhart *et al*, 2007; Haber & Barnhart, 2008).

Evaluating precision, or agreement, involves comparing multiple sets of test results which have measured the same quantity. Methods for quantifying the variability in test results are discussed in the following 2 sections. The same procedures that are used for measuring **precision** can be used to measure **agreement** between 2 different tests applied to the same sample.

#### 5.2.5 Measuring precision and agreement for tests with quantitative outcomes

Some commonly used techniques for quantifying variability, or for expressing results of comparisons between pairs of test results are:

- coefficient of variation
- Pearson correlation coefficient
- concordance correlation coefficient
- limits of agreement plots
- ICC

The **coefficient of variation** (CV) is computed as:

$$CV = \frac{\sigma}{\mu} \tag{Eq 5.1}$$

where  $\sigma$  is the standard deviation among test results on the same sample and  $\mu$  is the average of the test results. The CV for a given sample can be computed based on any number of repeat runs of the same test and then these values can be averaged over samples to compute an overall estimate of the CV (see Example 5.1).

A Pearson **correlation coefficient** measures the degree to which one set of test results (measured on a continuous scale) varies (linearly) with a second set. However, it does not directly compare the values obtained (it ignores the scales of the 2 sets of results) and for this reason, it is much less useful than a concordance correlation coefficient for comparing 2 sets of test results (see Example 5.1) and we do not recommend its use.

As with a Pearson correlation coefficient, a **concordance correlation coefficient (CCC)** (Lin, 1989; Lin, 2000) can be used to compare 2 sets of test results (eg results from 2 laboratories),

**Example 5.1 Measuring precision—quantitative test results**

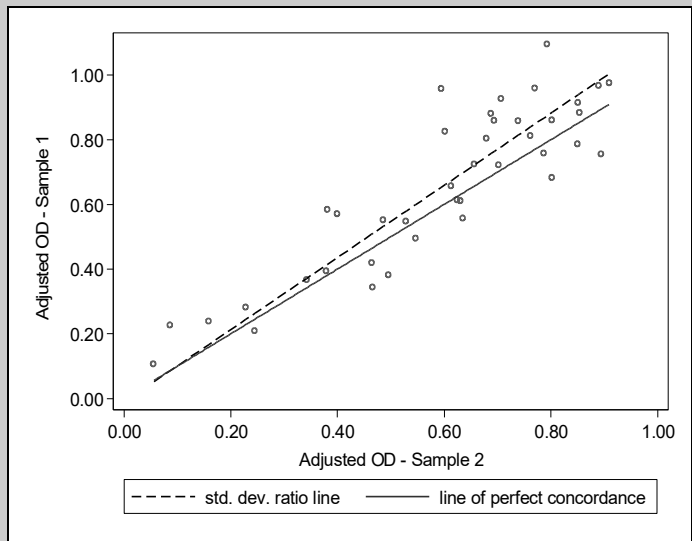
data = elisa\_repeat

A set of 40 individual cow milk samples was tested for parasite antibodies 6 times using an indirect microtitre ELISA based on a crude *Ostertagia ostertagi* antigen. Both raw and adjusted optical density (OD) values are recorded in the dataset (see Chapter 31 for description of the adjustment method). The results were used to evaluate the precision (repeatability) of the test.

The CV for each sample was computed based on the 6 replicate values and then averaged across the 40 samples. The mean CV values were 0.155 and 0.126, for the raw and adjusted values respectively.

Clearly,

adjustment of the values removed some of the plate-to-plate variability.



**Fig. 5.2 Concordance correlation plot**

Pearson correlation (not recommended) was used to compare values from replicates 1 and 2. The correlation was 0.937 for the raw values and 0.890 for the adjusted values.

Comparing replicates 1 and 2, the CCC was 0.762 for the raw values and 0.858 for the adjusted values, suggesting much better agreement between the 2 sets (replicates) of adjusted values (than between the 2 sets of raw values). (Note the lower values of the CCC compared to the Pearson correlation reflect the inadequacy of the Pearson correlation to detect differences between the 2 sets of results.) Fig. 5.2 shows a CCC plot for the adjusted values.

**Note** Data must overlay the solid line for perfect concordance. The reduced major axis is the linear regression line through the observations.

There appears to be a greater level of disagreement between the 2 sets of values at high OD readings compared with low OD readings.

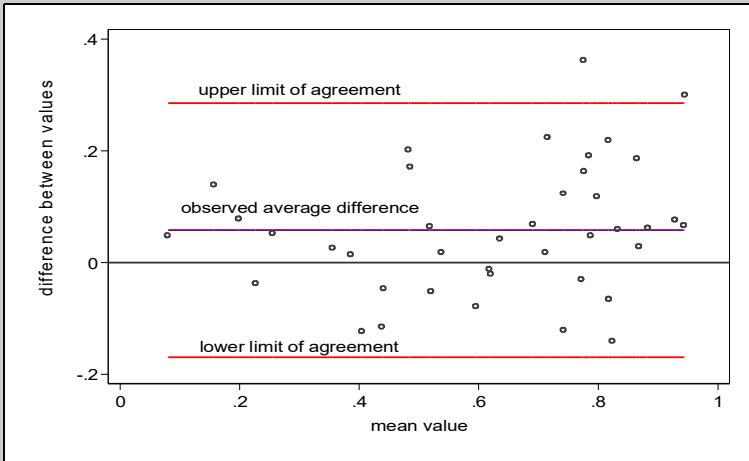
(continued on next page)

**Example 5.1** (continued)

data = elisa\_repeat

The limits of agreement plot for the same data is shown in Fig. 5.3. It indicates that most of the differences between the replicates fell in the range of -0.18 and +0.30 units.

All points would lie along the line  $y=0$  if there was perfect agreement between the 2 sets of results.



**Fig. 5.3** Limits of agreement plot

and it better reflects the level of agreement between the 2 sets of results than the Pearson correlation coefficient does. If 2 sets of continuous-scale test results agreed perfectly, a plot of one set against the other would produce a straight line at a 45° angle (the equality line). The *CCC* is computed from 3 parameters. The first, the location-shift parameter measures how far the data are (above or below) from the equality line. The second, the scale-shift parameter measures the difference between the slope for the sample data and the equality line (slope=1). (The product of the location-shift and scale-shift parameters is referred to as the **accuracy parameter**.) The third, the usual Pearson correlation coefficient, measures how tightly clustered the sample data are around the line (slope). The *CCC* is the product of the accuracy parameter and the Pearson correlation coefficient. A value of 1 for the *CCC* indicates perfect agreement. Example 5.1 shows a concordance correlation plot for 2 sets of ELISA results. The *CCC* has been generalised to deal with >2 sets of test results and to work with categorical data (Barnhart *et al*, 2002; King & Chinchilli, 2001) and to deal with the issue of repeated measurements (King *et al*, 2007).

A **limits of agreement plot** (also called a Bland-Altman plot) (Bland & Altman, 1986) plots the differences between the pairs of test results against their mean value. The mean ( $\mu_d$ ) and standard deviation of the differences ( $\sigma_d$ ) are computed and lines denoting the ‘limits of agreement’ are added to the plot at  $\mu_d \pm 1.96\sigma_d$ . These indicate the range of differences between the 2 sets of test results. This plot helps to determine if there is a systematic difference between the 2 sets of observations (*ie* mean difference < or > 0), and the range of errors (indicated by the spread of the points (de Vet, 2007)). The plot is also useful to determine if the level of disagreement between the 2 sets of results varies with the mean value of the substance being measured and can be used to identify the presence of outlying observations. A limits of

agreement plot is presented in Fig. 5.3.

**Reliability** is not strictly speaking a measure of precision because it relates the variability of a test result to the amount of variation among individuals (McDowell & Newell, 1996). Nevertheless, it is a term commonly encountered in the clinical epidemiology literature. Reliability is most commonly measured using the intraclass correlation coefficient (*ICC*) which is described in more detail in Section 20.3.3. In the context of diagnostic test evaluation, the *ICC* relates the amount of variability among individuals to the total variability which consists of variability among individuals plus variability among measurements within an individual (de Vet *et al.*, 2006).

$$ICC = \frac{\text{variability among individuals}}{\text{variability among individuals} + \text{measurement error}}$$

Alternatively, it can be viewed as 1-the proportion of variance due to measurement error. If a test is imprecise (much measurement error), the reliability will be low. See de Vet HCW *et al.* (2006) for a discussion on the use of agreement and reliability measures.

**5.2.6 Measuring precision and agreement for tests with a qualitative outcome**

All of the above procedures are useful if the quantity of interest is measured on a continuous scale. If the test results are categorical (dichotomous or multiple categories), a kappa (also called Cohen’s kappa) (Cohen, 1960) statistic can be used to measure the level of agreement between 2 (or more) sets of test results. Obviously, the assessments must be carried out independently of each other using the same set of outcome categories. The data layout for assessing agreement is shown in Table 5.1 for a 2X2 table (larger ‘square’ tables are also used).

**Table 5.1 Layout for comparing results from 2 qualitative (dichotomous) tests**

	Test 2 positive	Test 2 negative	Total
Test 1 positive	$n_{11}$	$n_{12}$	$n_{1.}$
Test 1 negative	$n_{21}$	$n_{22}$	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n$

**5.2.7 Kappa**

In assessing how well the 2 tests agree, we are not seeking answers relative to a gold standard (Section 5.3.1) as this might not exist, but rather whether the results of 2 tests agree with each other. Obviously, there will always be some agreement due to chance, and this must be considered in the analysis. For example, if one test was positive in 30% of subjects and the other test was positive in 40%, both would be expected to be positive in  $0.4*0.3=0.12$  or 12% of subjects by chance alone. So, the important question is: what is the level, or extent, of agreement beyond what would have been expected by chance? This question is answered by a statistic called Cohen’s kappa ( $\kappa$ ). We can calculate the essential elements of  $\kappa$  as follows:

- observed agreement =  $(n_{11} + n_{22})/n$
- expected agreement (chance) =  $[(n_{1.} * n_{.1})/n + (n_{2.} * n_{.2})/n]/n$
- actual agreement beyond chance = observed - expected
- potential agreement beyond chance = (1 - expected)

- $\kappa$  = actual agreement beyond chance/potential agreement beyond chance.

A formula for calculating  $\kappa$  directly is:

$$\kappa = 2 \frac{(n_{11}n_{22} - n_{12}n_{21})}{n_{1.}n_{2.} + n_{2.}n_{1.}} \tag{Eq 5.2}$$

Procedures for computing the standard error, confidence intervals and tests of significance for  $\kappa$  are available elsewhere (Reichenheim, 2004).

Common interpretations of  $\kappa$ , when applied to a test that is subjective in nature (eg identifying lesions on an X-ray), are shown below (Landis & Koch, 1977). One would expect to apply a more stringent interpretation when comparing 2 reasonably objective tests (eg virus isolation and PCR).

$\leq 0$	poor agreement
0.01 to 0.2	slight agreement
0.21 to 0.4	fair agreement
0.41 to 0.6	moderate agreement
0.61 to 0.8	substantial agreement
0.81 to 1.0	almost perfect agreement

Example 5.2 shows the computation of  $\kappa$  for assessing agreement between indirect fluorescent antibody test (IFAT) results for infectious salmon anemia (ISA) when the test was performed in 2 different laboratories (Nerette *et al*, 2005b).

### 5.2.8 Factors affecting kappa

It has been well established that bias (tendency of one rater to assign more positive test results than the other rater) and the prevalence of the underlying condition both affect  $\kappa$  (Cook, 2007; Nam, 2007; Sargeant & Martin, 1998). Alternative approaches to measuring agreement have been suggested and these include: maximum kappa (Feinstein & Cicchetti, 1990), Yule’s Y (Spitznagel & Helzer, 1985), indices of positive and negative agreement (Cicchetti & Feinstein, 1990), and a prevalence and bias adjusted kappa (PABAK—also called the *S* coefficient) (Byrt *et al*, 1993; Thomsen & Baadsgaard, 2006) and conditional relative odds ratio (Suzuki, 2006). However, in general, these have not been widely adopted so it is important to consider the role of bias and prevalence on  $\kappa$ .

**Bias** Before quantifying the level of agreement, we need to determine if the 2 tests are classifying approximately the same proportion of individuals as positive. (If one test produces more positive test results than the other, there is not much point in proceeding to evaluate agreement.) We compare the proportion positive to each test (*ie*  $p_1$  and  $p_2$ , where  $p_1$  and  $p_2$  represent the proportion positive to tests 1 and 2, respectively) using the **McNemar’s  $\chi^2$**  test for paired data (Lachenbruch, 2007; McNemar, 1947) or an exact binomial test for correlated proportions (formula not shown).

$$\text{McNemar's } \chi^2 = (n_{12} - n_{21})^2 / (n_{12} + n_{21}) \tag{Eq 5.3}$$

A non-significant test indicates that there is little evidence that the 2 proportions positive differ. If significant, this test suggests a serious disagreement between the tests and thus the detailed assessment of agreement could be of little value.



**Example 5.2 Agreement among dichotomous test results**

data = ISA\_test

Kidney samples from 291 salmon were split with one-half being sent to each of 2 laboratories where an IFAT test was run on the sample. IFAT results were expressed as 0 (negative) or 1+, 2+, 3+, or 4+. They were subsequently dichotomised so that all scores of 1+ and higher were considered positive. The data were:

	IFAT 2 positive	IFAT 2 negative	Total
IFAT 1 positive	19	10	29
IFAT 1 negative	6	256	262
Total	25	266	291

McNemar's  $\chi^2$  test had the value 1.00 (P=0.317; the binomial P-value was 0.45) indicating that there is little evidence that the 2 laboratories found different proportions positive.

observed agreement = 0.945

expected agreement = 0.832

$\kappa = 0.674$

SE( $\kappa$ )<sup>a</sup> = 0.0762

95% CI of  $\kappa = 0.524$  to 0.823

Thus, the level of agreement appears substantial. However, the CI is wide, reflecting considerable uncertainty about the estimate.

<sup>a</sup> There are a number of formulae for the SE; the one used here does not assume independence of observations

**Prevalence** As noted, the prevalence of the condition being diagnosed affects  $\kappa$ . Two tests (or 2 raters) will have a higher  $\kappa$  value if the prevalence of the underlying condition is moderate (~0.5) than if it is very high or very low. The relationship between prevalence and  $\kappa$  is complex, and depends on the distribution of difficult-to-classify individuals in the population. However, in general, the influence of prevalence is only substantial at very high and very low prevalence values. A much more detailed review of this issue, and the conclusion that we should not be concerned about the effect of prevalence on  $\kappa$ , has recently been published (Vach, 2005).

**5.2.9 Multiple raters (tests)**

Kappa can be extended to situations in which there are more than 2 raters (or tests). In this instance, there is no assumption about the uniqueness of the raters, so an individual may be evaluated by different numbers of raters or by the same number of raters but with different individuals doing the rating. (However, a balanced study in which the same raters evaluate all individuals will provide the most meaningful results.) This same approach can be used when there are only 2 raters, but the identity of those raters differs across subjects. Details of these methods are covered in Fleiss *et al* (2003).

When data from multiple raters are available, an alternative to computing  $\kappa$  is to model the probability of a positive test result using a multilevel model (see Chapter 22) with the rater as a random effect (Woodard *et al*, 2007). This analysis focuses on factors that affect the probability of a positive test result, but the estimate of the between-rater variance provides some insight into the level of agreement.

### 5.2.10 Weighted kappa

For tests measured on an ordinal scale, computation of the usual  $\kappa$  assumes that any pair of test results which are not in perfect agreement are considered to be in disagreement. However, if a test result is scored on a 5-point scale, a pair of tests with scores of 5 and 4, respectively should be considered in ‘less disagreement’ than a pair of scores of 5 and 1. Partial agreement can be taken into account using a weighted  $\kappa$  in which pairs of test results that are close are considered to be in partial agreement (through a weight matrix which specifies how much agreement should be assigned to them). A weighted  $\kappa$  is sensitive to the number of categories used (Brenner & Kliebsch, 1996) and to the choice of weights (Graham & Jackson, 1993). Confidence intervals can be computed using bootstrap methods (Reichenheim, 2004) and an exact test of statistical significance is available (Brusco *et al*, 2007) (although we are usually more interested in the magnitude of  $\kappa$  than in its statistical significance).

Example 5.3 shows the data layout and the results of an unweighted and weighted  $\kappa$  for comparing 2 sets of IFAT results for the ISA virus in salmon. It has been reported that computing an intraclass correlation coefficient may be superior to the use of a weighted kappa when dealing with ordinal response categories (Maclure & Willett, 1987).

## 5.3 THE ABILITY OF A TEST TO DETECT DISEASE OR HEALTH

The 2 key characteristics we estimate are the ability of a test to detect diseased animals correctly (its **sensitivity**), and to give the correct answer if the animal in question is not diseased (its **specificity**). For pedagogical purposes, we will assume that animals are the units of interest (the principles apply to other levels of aggregation). Further, we will assume that a specific ‘disease’ is the outcome although other conditions such as pregnancy, premature removal (culling), having a specified antibody titre, or infection status could be substituted in a particular instance. To initiate this discussion, it is simplest to assume that the test we are evaluating gives only dichotomous answers—positive or negative. This might be a bacterial culture in which the organism is either present or absent, or a survey question about whether or not a dairy farmer uses a milking machine with automatic take-offs. In reality, many test results provide a continuum of responses and a certain level of response (colour, test result relative to background signal, level of enzyme activity, endpoint titre *etc*) is selected such that, at or beyond that level, the test result is deemed to be positive.

### 5.3.1 The gold standard

A gold standard is a test or procedure that is absolutely accurate. It diagnoses all of the specific disease that exists and misdiagnoses none. For example, if we had a definitive test for feline leukemia virus infection that correctly identified all feline leukemia-virus infected cats as positive and gave negative results in all non-infected cats, the test would be considered a gold standard. In reality, there are very few true gold standards. Partly this is related to imperfections in the test itself, but a good portion of the error is due to biological variability. Animals do not immediately become ‘diseased’, even subclinically, when exposed to an infectious, toxic, physical or metabolic agent. Usually, a period of time will pass before the animal responds in a manner that produces a detectable or meaningful change. The time period for an animal’s response to cross the threshold and be considered ‘diseased’ varies from animal to animal.

**Example 5.3 Agreement among ordinal test results**  
 data = ISA\_test

The data described in Example 5.2 were used except the original ordinal data were retained (5-point scale).

IFAT 1	IFAT 2				
	Neg	+	++	+++	++++
Neg	256	5	0	1	0
+	8	2	0	2	0
++	2	1	0	4	0
+++	0	0	2	2	0
++++	0	0	0	3	3

An unweighted kappa (which assumes that all test results which were not identical were in disagreement) and a weighted kappa in which test results were:

- identical: weighted as complete agreement
- 1 level apart: weighted as 70% agreement
- 2 levels apart: weighted as 30% agreement
- >2 levels apart: weighted as complete disagreement.

	Kappa	SE	95% CI	
			Lower	Upper
Unweighted	0.450	0.037	0.318	0.569
Weighted	0.693	0.046	0.570	0.793

The weighted kappa suggests substantial agreement and is probably a better reflection of the agreement between the 2 sets of tests than the unweighted kappa.

Traditionally, in order to assess a new test, we required a gold standard. However, a variety of approaches for evaluating diagnostic tests are discussed in Section 5.7

**5.3.2 Sensitivity and specificity**

The concepts of sensitivity and specificity are often easier to understand through the use of a 2X2 table, displaying disease and test results in a sample of animals.

**Table 5.2 Data layout for test evaluation**

	Test positive (T+)	Test negative (T-)	Total
Disease positive (D+)	a (true positive)	b (false negative)	m <sub>1</sub>
Disease negative (D-)	c (false positive)	d (true negative)	m <sub>0</sub>
Total	n <sub>1</sub>	n <sub>0</sub>	n

The sensitivity of a test (*Se*) is the proportion of diseased animals that test positive. It is described statistically as the conditional probability of testing positive given that the animal is diseased [p(T+|D+)], and is measured by:

$$Se = \frac{a}{a+b} = \frac{a}{m_1} \quad \text{Eq 5.4}$$

The specificity of a test ( $Sp$ ) is the proportion of non-diseased animals that test negative. It is described statistically as the conditional probability of testing negative given that the animal does not have the disease of interest [ $p(T|D^-)$ ] and is measured by:

$$Sp = \frac{d}{c+d} = \frac{d}{m_0} \quad \text{Eq 5.5}$$

For future purposes, we will denote the false positive fraction ( $FPF$ ) as  $1-Sp$  and the false negative fraction ( $FNF$ ) as  $1-Se$ . From a practical perspective, if you want to confirm a disease, you would use a test with a high  $Sp$  because there are few false positives. Conversely, if you want to rule out a disease, you would use a test with a high  $Se$  because there are few false negatives. Confidence intervals for  $Se$ ,  $Sp$ ,  $FPF$  and  $FNF$  can be obtained using procedures applicable for estimating the confidence interval of a proportion (see Section 4.10). Estimates of  $Se$  and  $Sp$  are specific for a given population and may vary across source populations. Methods for estimating  $Se$  and  $Sp$  are covered in Section 5.7 and 5.8 and factors that might affect the  $Se$  and  $Sp$  are discussed in Section 5.9.

The estimation of  $Se$  and  $Sp$  of an indirect ELISA test for detecting bovine fetuses persistently infected (PI) with the bovine virus diarrhea (BVD) virus is shown in Example 5.4. A blood sample is taken from the cow in late lactation and tested for antibodies to the virus. If antibodies are present at a high level, the fetus is deemed to be persistently infected with the BVD virus.

### 5.3.3 True and apparent prevalence

Two other terms are important descriptors of the tested subgroup. One denotes the actual level of disease that is present. In screening-test jargon, this is called the **true prevalence** ( $P$ ); in clinical epidemiology, this is referred to as **prior prevalence**, or **pre-test prevalence**.  $P$  is a useful piece of information to include in our discussion of test evaluation because it will affect the interpretation of the test result. In Example 5.4,  $P=p(D^+)=m_1/n=233/1673=0.139$  or 13.9%.

In contrast to the 'true' state, unless our test is perfect, the test results will only provide an estimate of the true prevalence and, in screening-test jargon, this is called the **apparent prevalence** ( $AP$ ). In Example 5.4,  $AP=p(T^+)=n_1/n=800/1673=0.478$  or 47.8%. In clinical epidemiology, this might be referred to as a **post-test prevalence**. In general,  $AP$  can be computed as:

$$AP = p(T^+) = P * Se + (1 - P)(1 - Sp) \quad \text{Eq. 5.6}$$

### 5.3.4 Estimating true prevalence from apparent prevalence

If the  $Se$  and  $Sp$  of a test are known, the true prevalence of disease in a population is estimated by Rogan & Gladen (1978):

$$p(D^+) = \frac{AP - (1 - Sp)}{1 - [(1 - Sp) + (1 - Se)]} = \frac{AP + Sp - 1}{Se + Sp - 1} \quad \text{Eq 5.7}$$

For example, if  $AP=0.150$  and  $Se=0.363$ ,  $Sp=0.876$ , then our estimate of true prevalence is

**Example 5.4 Sensitivity, specificity and predictive values**

data = bvd\_test

The data used for this example came from a study done to evaluate an ELISA test for the diagnosis of bovine fetuses persistently infected (PI) with BVD virus. See Chapter 31 for a more complete description of this dataset. The test was designed to work on both milk and blood samples, but the data used here relate only to the blood sample results. The mean optical density was 0.92 units. Thus, for this example a fetus was deemed to be test positive if the optical density of the blood test was greater than 0.92 units. (This is not an optimal cutpoint for this test, but is used for illustration.)

		T+	T-	
<b>PI+</b>	(D+)	178	55	233
<b>PI-</b>	(D-)	622	818	1440
		800	873	1673

For purposes of description, the 178 animals are called true positives, the 622 are false positives, the 55 are false negatives and the 818 are true negatives. We will assume here that the study subjects were obtained using a simple random sample.

In this example,

- $Se = 178/233 = 76.4\%$     95% CI = (70.4% , 81.7%)
- $Sp = 818/1440 = 56.8\%$     95% CI = (54.2% , 59.4%)
- $FNF = 1-0.764 = 23.6\%$
- $FPF = 1-0.568 = 43.2\%$
- $P = 233/1673 = 13.9\%$
- $AP = 800/1673 = 47.8\%$
- $PV+ = 178/800 = 22.3\%$     95% CI = (19.4% to 25.3%)
- $PV- = 818/873 = 93.7\%$     95% CI = (91.9% to 95.2%)

**Note** The confidence intervals are exact based on the binomial distribution but do not take into account the clustering of observations within herds.

0.109 or 10.9%. It is possible that some combinations of  $Se$ ,  $Sp$  and  $AP$  result in estimates of true prevalence outside its allowed range (0-1). This indicates that one or both of the  $Se$  and  $Sp$  estimates used are not applicable for the population being studied.

**5.4 PREDICTIVE VALUES**

The  $Se$  and  $Sp$  are characteristics of the test. However, these terms do not tell us directly how useful the test might be when applied to animals of unknown disease status. Once we have decided to use a test, we want to know the probability that the animal has or does not have the disease in question, depending on whether it tests positive or negative. These probabilities are called **predictive values** and these change with different populations of animals tested with the same test because they are driven by the true prevalence of disease in the study population as well as by the test characteristics. In this discussion, we assume the group of subjects being tested is homogeneous with respect to the true prevalence of disease. If not, then the covariates that affect disease risk should be identified and separate estimates made for each subpopulation.

### 5.4.1 Predictive value positive

With data as shown in Table 5.2, the predictive value of a positive test ( $PV+$ ) is the probability that given a positive test, the animal actually has the disease; this might be represented as  $p(D+|T+)$  or  $a/n_1$  in Table 5.2. The predictive value of a positive test can generally be estimated using the following formula:

$$PV+ = \frac{p(D+) * Se}{p(D+) * Se + p(D-) * (1 - Sp)} \quad \text{Eq 5.8}$$

This formula explicitly shows how the true prevalence of disease in the tested group affects the  $PV+$ .

### 5.4.2 Predictive value negative

In a similar manner, the  $PV$  of a negative test ( $PV-$ ) is the probability that given a negative test, the animal does not have the disease (ie  $p(D-|T-)$ ). In Table 5.2 this is  $PV- = d/n_0$ . The predictive value of a negative test result can be estimated using the following formula:

$$PV- = \frac{p(D-) * Sp}{p(D-) * Sp + p(D+) * (1 - Se)} \quad \text{Eq 5.9}$$

Estimates of  $PV+$  and  $PV-$  are shown in Example 5.4. **Note** These values represent the predictive values given the  $P$  observed in the study population.

Because we are more often interested in the ‘disease’ side of the question, there is a measure of the probability that an animal that tests negatively is actually diseased. It is called the positive predictive value of a negative test or  $PPV- = b/n_0$  or  $1 - (PV-)$ .

### 5.4.3 Effect of prevalence on predictive values

As noted above, the predictive values of the test depend on the sensitivity and specificity of the test and the prevalence of the disease in the population in which it is used. Consequently, predictive values are not good measures of a test’s performance (because they vary from population to population). Example 5.5 shows how dramatically predictive values can change as the prevalence of a disease varies from 50% to 1%.

Computing confidence intervals (CI) for  $PVs$  is not straightforward. The CI at the observed  $P$  can be computed as a CI for a binomial proportion (see Section 4.10) given the observed sample size. In situations in which the  $PV+$  or  $PV-$  approaches 1 (often the  $PV-$  approaches 1 when  $P$  is small) exact methods of computing CI for binomial proportions, or other methods of dealing with the problem that the CI that may extend below 0 or above 1 should be employed (Mercaldo *et al*, 2007). In order to estimate  $PVs$  for values of  $P$  other than that observed in the data, the uncertainty about the estimates of the  $Se$  and  $Sp$  as well as the estimate of  $P$  need to be taken into account (see Zou (2004) for a discussion of the problem and one approach to computing these CIs).

### 5.4.4 Increasing the predictive value of a positive test

One way to increase the predictive value of a positive test is to use the test on animals where the

**Example 5.5 Effect of prevalence on predictive values**

data = bvd\_test

In order to examine the impact of a change in  $P$  on the outcome of a test, we will use the values of  $Se$  and  $Sp$  from Example 5.4 and specify 3 scenarios where the true prevalence varies from 50% to 5%, and then to 1%. For pedagogical purposes, we demonstrate the calculations for the 50% prevalence scenario in a 2X2 table. A simple way to proceed to obtain these results is to construct a fictitious population of 1,000 animals with 500 being ‘diseased’ (ie PI+) and 500 being PI- based on the true prevalence of 50%. Then, we calculate 76.4% ( $Se$ ) of 500 and fill in the 382 true positives. Finally, we calculate 56.8% ( $Sp$ ) of 500, fill in the 284 true negatives, and complete the table.

	Test +	Test -	
PI+	382	118	500
PI-	216	284	500
	598	402	1000

From these data:

$PV+ = 382/598 = 63.9\%$       The probability that a cow testing positive will truly have a PI+ calf is 63.9%

$PV- = 284/402 = 70.6\%$ .      The probability that a cow testing negative will truly have a PI- calf is 70.7%

Comparable values if the prevalence is 5% or 1% are:

Prevalence (%)	PV+ (%)	PV- (%)
5	8.5	97.9
1	1.8	99.6

As you can see, the  $PV+$  drops off rapidly as  $P$  falls, but the  $PV-$  rises.

prevalence in the population being tested is relatively high. Thus, in a screening programme designed to ascertain if a disease is present, we often might slant our testing toward animals that are likely to have the disease in question. Hence, testing culled animals, or animals with a particular history, is a useful way of increasing the pre-test (prior) probability of disease.

A second way to increase  $PV+$  is to use a more specific test (with the same or higher  $Se$ ), or change the cutpoint of the current test to increase the  $Sp$  (but this would decrease the  $Se$  somewhat also). As  $Sp$  increases,  $PV+$  increases because the number of false positives approaches zero. A third, and very common way to increase  $PV+$  is to use more than one test. Here the result depends on the method of interpretation as well as the individual test characteristics.

**5.5 INTERPRETING TEST RESULTS MEASURED ON A CONTINUOUS SCALE**

For many tests, the substance being evaluated (eg urea in milk, serum calcium, liver enzymes) is measured on a continuous scale or with semi-quantitative (ordinal) results. Predictive probabilities associated with these test results can be used directly to estimate the prevalence of disease in a population (Choi *et al*, 2006). However, to interpret the result at an individual level, we need to select a **cutpoint** (also called **cut-off** or **threshold**) to determine what level of result

indicates a positive test result. This is also true when interpreting serologic titres.

In reality, there is often an overlap in the distribution of the substance being measured between healthy and diseased animals and we usually select a cutpoint that optimises the  $Se$  and  $Sp$  of the test. The dilemma is depicted in Fig. 5.4. As will be demonstrated (Section 5.5.3), it is often useful to use the actual result when assessing the health status of the tested subject(s).

### 5.5.1 Selecting a cutpoint

If there is any overlap in the test values for D+ and D- animals, whatever cutpoint we choose will result in both false positive and false negative test results (eg Fig. 5.4). For the BVD data, the distributions of optical density ( $OD$ ) values in the PI+ and PI- calves overlap considerably. Because of this overlap, if we raise the cutpoint, the  $Sp$  will increase (false positives decrease) and the  $Se$  will decrease (more false negatives). Lowering the cutpoint has the opposite effect. Thus, the choice of cutpoint will depend on the relative seriousness of either a false negative or a false positive test result.

If one has to choose among multiple cutpoints, graphical procedures such as **receiver operating characteristic curves** (ROC—described below) or a **sensitivity-specificity plot** (also called a **2-graph ROC plot**) might be used to help choose an optimal cutpoint. Alternatively, it is possible to use the actual test result value by computing likelihood ratios (see Section 5.5.3) and avoid having to select a specific cutpoint.

A sensitivity-specificity plot (Reichenheim, 2002) shows how the  $Se$  and  $Sp$  of a test changes as the cutpoint is moved through the possible range of values (Fig. 5.5). It can be used to identify where the 2 values are equal, but this is not necessarily the best cutpoint. Depending on the cost of false positive and false negative test results, it may be important to choose a cutpoint which results in high  $Se$  (and consequently relatively low  $Sp$ ) or vice versa. As can be seen in Fig 5.5, obtaining a  $Se$  much greater than 70% for the BVD test entails accepting quite a low  $Sp$ .

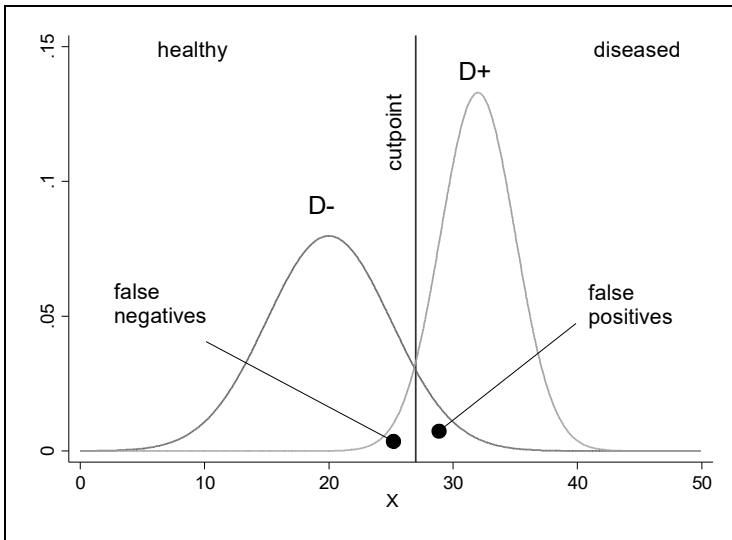


Fig. 5.4 Overlap between healthy and diseased animals

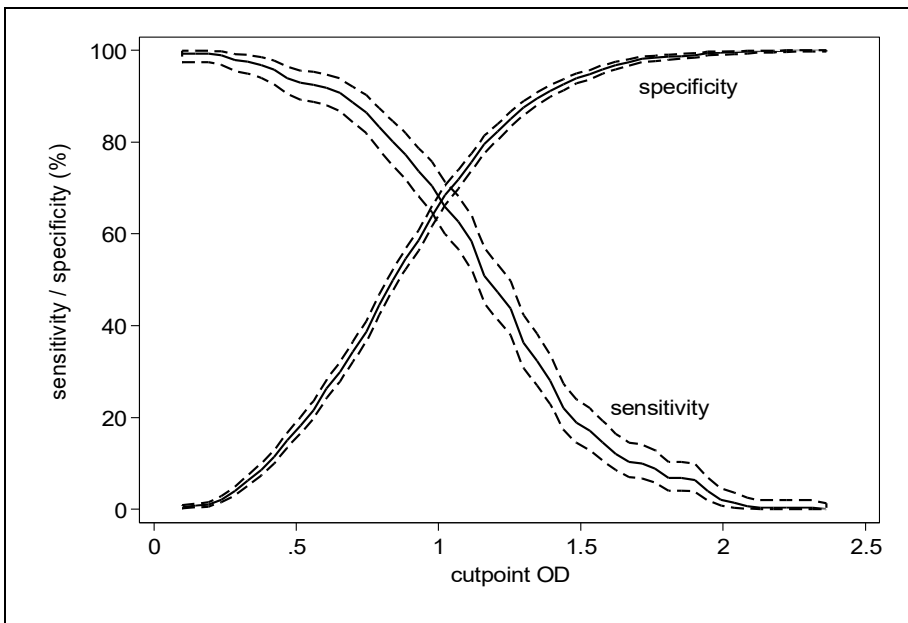


### 5.5.2 Receiver operating characteristic curves

A receiver operating characteristic (ROC) curve is a plot of the  $Se$  of a test versus the false positive rate ( $1-Sp$ ) computed at a number of different cutpoints to select the optimum cutpoint for distinguishing between diseased and non-diseased animals (Greiner *et al*, 2000). The 45° line in Fig. 5.6 represents a test with discriminating ability that is no better than chance alone. The closer the ROC curve gets to the top-left corner of the graph, the better the ability of the test to discriminate between diseased and non-diseased animals. (The top-left corner represents a test with a  $Se$  of 100% and a  $Sp$  of 100%).

Use of an ROC curve has the advantage over a ‘one cutpoint value’ for determining  $Se$  and  $Sp$  in that it describes the overall ability of the test to discriminate diseased from non-diseased animals over a range of cutpoints. The area under the ROC curve (AUC) can be interpreted as the probability that a randomly selected  $D+$  individual has a greater test value (*eg* optical density) than a randomly selected  $D-$  individual (again assuming the distribution of the test results in the  $D+$  group is higher than that in the  $D-$  group). Multiple approaches to estimating the SE of the AUC are available and have been reviewed (Faraggi & Reiser, 2002; Hajian-Tilaki & Hanley, 2002). ROC analysis can also be used to compare 2 (or more) tests based on the AUC, see Pepe (2003) for details.

Assuming equal costs of false negative and false positive test results, the optimal cutpoint is that with  $Se+Sp$  at a maximum, and this occurs where the curve gets closest to the top left corner of the graph (or alternatively, the farthest away from the 45° line). Depending on the seriousness of false negative versus false positive results, one might want to emphasise test results in one particular region of the ROC curve such as an area that constrains  $Se$  (or  $Sp$ ) within defined limits. This is referred to as the partial AUC (Walter, 2005b).



**Fig. 5.5 Sensitivity-specificity plot of the BVD test data (with 95% confidence intervals (dashed lines))**

Both parametric and non-parametric ROC curves can be generated. A non-parametric curve simply plots the  $Se$  and  $(1-Sp)$  using each of the observed values of the test result as a cutpoint. A parametric ROC curve provides a smoothed estimate by assuming that the latent variables representing the  $Se$  and  $(1-Sp)$  at various cutpoints follow a specified distribution (usually binormal). Example 5.6 shows both parametric and non-parametric ROC curves for the `bvd_test` data. Recently, a semi-parametric ROC curve has been proposed (Wan & Zhang, 2007).

### 5.5.3 Likelihood ratios

A likelihood ratio ( $LR$ ) is the ratio of the probability of a given test result among diseased individuals to the probability of that test result among non-diseased individuals. Consequently, for a test with a dichotomous test result, there are 2  $LR$ s: one for a positive test results ( $LR+$ ) and one for a negative test result ( $LR-$ ). The  $LR+$  is the ratio of the post-test odds of disease divided by the pre-test odds. Recall that, in general, an odds is  $P/(1-P)$  so an  $LR$  of a positive test result is the odds of disease given the test result divided by the pre-test odds:

$$LR+ = \frac{PV+/(1-PV+)}{P/(1-P)} = \frac{Se}{1-Sp} \quad \text{Eq 5.10}$$

where  $P$ =prevalence or  $p(D+)$  in the group being tested. Consequently,  $LR$ s reflect how our view changes of how likely disease is when we get the test result.

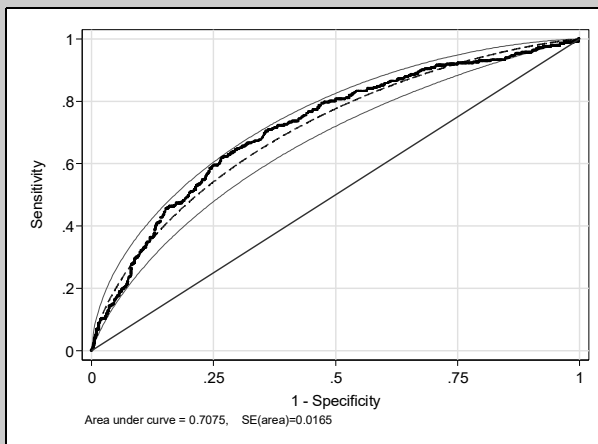
For tests with continuous outcomes there are 3 possible  $LR$ s (Choi, 1998; Gardner & Greiner, 2006):

- test value specific

#### Example 5.6 ROC curves

data = `bvd_test`

Fig. 5.6 shows both non-parametric (points) and parametric (dashed line) ROC curves along with 95% CI curves for the parametric ROC.



**Fig. 5.6 Parametric and non-parametric ROC curves**

- cutpoint specific, and
- category specific.

A test value specific *LR* is the ratio of the probabilities of an exact test result in diseased and non-diseased individuals. Because of sample size limitations it is not usually computed. However, it can be estimated by determining the tangent to the ROC curve at that test value (Choi, 1998).

**5.5.4 Cutpoint specific *LR***

A cutpoint specific *LR* ( $LR_{cp}$ ) at a selected cutpoint is the ratio of the probabilities of test results above the cutpoint in D+ individuals to that in D- individuals. It can be written as :

$$LR_{cp+} = \frac{Se_{cp}}{1 - Sp_{cp}} \tag{Eq 5.11}$$

where cp denotes the cutpoint at or above which the test is considered positive. In this context, the  $LR+$  can be viewed as the probability of a diseased individual having a test result above the cutpoint relative to the probability of the same result in a non-diseased subject. The  $LR_{cp+}$  can be estimated as the slope of the line from the origin to the cutpoint on an ROC curve (Choi, 1998).

The *LR* for a negative test result ( $LR-$ ) at a given cutpoint is the ratio  $(1-Se)/Sp$ . It denotes the probability of the negative result from a diseased subject relative to that of a non-diseased subject. Examples of *LRs* at various cutpoints are shown in Example 5.7.

The *LR* makes use of the actual test result (as opposed to just being positive) and gives a quantitative estimate of the increased probability of disease given the observed result. For example, at the cutpoint 1.1, the  $LR+$  is 2.31, meaning that a cow that tests positive at this cutpoint (*ie* a test result  $\geq 1.1$ ) is 2.3 times more likely to have a PI+ calf than you thought it was prior to testing. **Note** Technically, we should state that the odds, rather than the probability, of the disease has gone up 2.3 times but if the disease is rare, then  $odds \approx probability$ . This approach makes use of the fact that, in general, the *LR* increases as the strength of the response (test result) increases.

**5.5.5 Category specific *LR***

Often researchers in a diagnostic setting prefer to calculate *LRs* based on the **category-specific** result ( $LR_{cat}$ ) as opposed to the cumulative distributions (Giard and Hermans, 1996).

Here the *LR* is:

$$LR_{cat} = \frac{P(\text{result} | D+)}{P(\text{result} | D-)} \tag{Eq 5.12}$$

The  $LR_{cat}$  can be estimated as the slope of the line joining 2 points of an ROC curve that represent the boundaries of the category.

Regardless of how they are computed, *LRs* are useful because they combine information on both sensitivity and specificity and they allow the determination of post-test from pre-test odds of disease as shown:

**Example 5.7 Likelihood ratios**

data = bvd\_test

Cutpoint-specific and category-specific likelihood ratios for the bvd\_test data

Optical density cutpoint	PI+ category (%)	Cumulative sensitivity (%)	LR cp+	PI- category (%)	Cumulative specificity (%)	LR cp-	LR cat
0	6.76	100.00		16.91	0.00		0.40
0.5	4.63	93.24	1.12	17.44	16.91	0.40	0.27
0.7	13.17	88.61	1.35	21.64	34.34	0.33	0.61
0.9	16.01	75.44	1.71	18.29	55.98	0.44	0.88
1.1	23.13	59.43	2.31	13.29	74.27	0.55	1.74
1.3	18.51	36.30	2.92	6.75	87.56	0.73	2.74
1.5	7.83	17.79	3.13	3.62	94.31	0.87	2.17
1.7	3.56	9.96	4.81	1.01	97.93	0.92	3.52
1.9	6.41	6.41	6.02	1.06	98.94	0.95	6.02

Categories are computed with the cutpoint shown as the left-hand end of the category (eg the category for cutpoint 1.5 is from 1.5 to 1.699). Results are based on 2,162 test results.

$$\text{post-test odds} = LR * \text{pre-test odds} \tag{Eq 5.13}$$

When interpreting the post-test odds, we need to be aware of whether the  $LR_{cp}$  or  $LR_{cat}$  is being used. The former gives the post-test odds for an animal testing positive at that level or higher, whereas the latter gives the post-test odds for animals testing positive in that specific category (or level) of test result. The process of computing the category-specific post-test probability is as follows, assuming that, prior to testing, you thought there was a 2% probability of the cow having a PI+ fetus and that the test OD was 1.77 ( $LR_{cat}=3.52$  (from Example 5.6)):

- convert the pre-test probability to pre-test odds  
pre-test odds =  $0.02/0.98=0.0204$
- multiply the pre-test odds by the likelihood ratio to get the post-test odds  
post-test odds =  $0.0204*3.52=0.0718$
- convert the post-test odds to a post-test probability  
post-test probability =  $0.0718/(1+0.0718)=0.067$

After obtaining a test result of 1.77, your estimate of the probability that the cow is carrying a PI+ fetus is 6.7%.

The variance of the  $\ln LR_{cp}$  is:

$$\text{var}(\ln LR_{cat}) = (1 - p(\text{result}|D+)) / a + (1 - p(\text{result}|D-)) / b \tag{Eq 5.14}$$

where  $a$  and  $b$  are the number of individuals with the result of interest in the  $D+$  and  $D-$  groups, respectively. A  $(1-\alpha)\%$  CI is:

$$LR_{cat} * \exp(\pm Z_{\alpha} \sqrt{\text{var}(\ln LR_{cat})}) \tag{Eq 5.15}$$

## 5.6 USING MULTIPLE TESTS

As stated, the use of multiple tests is an often-used approach to improve the overall diagnostic ability of the screening (or diagnostic) process.

### 5.6.1 Parallel and series interpretation

Using 2 tests represents the simplest extension of more than one test although the principles discussed below hold true for multiple tests. Suppose we have 2 different tests for detecting a disease. In Example 5.8, we use the results from the IFAT test for infectious salmon anemia ( $Se=0.784$ ,  $Sp=0.951$ ) and the polymerase chain reaction (PCR) test for the same disease ( $Se=0.926$ ,  $Sp=0.979$ ). If both tests are carried out, the results can be interpreted in one of 2 ways. With **series** interpretation, only animals that test positive to both tests are considered test positive. With **parallel** interpretation, animals that test positive to one test, the other test or both tests are considered test positive. Series interpretation increases  $Sp$  but decreases  $Se$ ; whereas parallel testing increases  $Se$  and decreases  $Sp$ .

Tests are considered **conditionally independent** if the probability of getting a given test result on one test does not depend on the result from the other test, given the disease status of the individual. For example, assume that you are dealing with a non-diseased individual. Two tests will be conditionally independent if the probability of a false positive on test #2 is the same in individuals that were  $T^-$  on test 1 and in those that were  $T^+$  on test 1. If tests are conditionally independent, the formulae for  $Se$  and  $Sp$  under parallel ( $Se_p$ ,  $Sp_p$ ) and series ( $Se_s$ ,  $Sp_s$ ) interpretation are:

$$Se_p = Se_1 + Se_2 - (Se_1 * Se_2) \tag{Eq. 5.16}$$

$$Sp_p = Sp_1 * Sp_2 \tag{Eq. 5.17}$$

$$Se_s = Se_1 * Se_2 \tag{Eq. 5.18}$$

$$Sp_s = Sp_1 + Sp_2 - (Sp_1 * Sp_2) \tag{Eq. 5.19}$$

**Note** If tests are going to be interpreted in series, it often makes sense to first test all animals with the test that is less expensive and/or more rapid, and then test all test positives with the second test. This is referred to as **sequential testing** and it provides the same results as simultaneous testing, but at lower cost, because only those subjects/samples positive to the first test are followed-up with the second test.

### 5.6.2 Correlated test results

Given the previous discussion on parallel and series interpretation, one might think that virtually 100%  $Se$  would be obtainable with 2-to-3 tests used in parallel, or 100%  $Sp$  with 3-to-4 tests used in series. However, Example 5.8 uses observed values, not ones we might expect assuming conditional independence of tests. The expected distributions of results, if the tests were independent, are shown in Table 5.3.

**Example 5.8 Multiple tests—series versus parallel interpretation**

data = ISA\_test

The data in this example are from the ISA\_test dataset. The tests we are using are the indirect fluorescent antibody test (IFAT) and the polymerase chain reaction (PCR) test, with clinical disease status (see dataset description Chapter 31) as the gold standard. The observed joint distributions of test results and virus presence are shown below along with the 4 possible test interpretation criteria.

	Number of fish by test-result category				Totals
<b>IFAT result</b>	+	+	0	0	
<b>PCR result</b>	+	0	+	0	
<b>Diseased fish</b>	134	4	29	9	176
<b>Non-diseased fish</b>	0	28	12	534	574
<b>Series interpretation</b>	+	0	0	0	
<b>Parallel interpretation</b>	+	+	+	0	

$Se$  of IFAT only =  $138/176 = 0.784$

$Sp$  of IFAT only =  $546/574 = 0.951$

$Se$  of PCR only =  $163/176 = 0.926$

$Sp$  of PCR only =  $562/574 = 0.979$

$Se$  of series interpretation =  $134/176 = 0.761$

$Se$  of parallel interpretation =  $(134+4+29)/176 = 0.949$

$Sp$  of series interpretation =  $(28+12+534)/574 = 1.000$

$Sp$  of parallel interpretation =  $534/574 = 0.930$

**Table 5.3 Expected Se and Sp levels with combined tests for ISA assuming conditional independence (data from Example 5.8)**

Interpretation	Sensitivity		Specificity	
	Expected	Observed	Expected	Observed
<b>Parallel</b>	$0.784+0.926 -$ $0.784*0.926=0.984$	0.949	$0.951*0.979=0.931$	0.930
<b>Series</b>	$0.784*0.926=0.726$	0.761	$0.951+0.979 -$ $0.979*0.951=0.999$	1.000

The expected  $Se$  for parallel interpretation is slightly higher than observed and slightly lower than observed for series interpretation. The expected and observed values for  $Sp$  are virtually identical. Note that **conditional independence** assumes that, in  $D+$  animals, the probability of a positive test result to test 2 is the same in samples that test negative to test 1 as it is in those that test positive to test 1. A similar assumption exists in  $D-$  individuals. More likely, and as observed with these data, especially if the tests are biologically related (eg both antibody tests), if test 1 is negative, the result on test 2 is more likely to be negative than if test 1 was positive. In this instance, we would describe the test results as **conditionally dependent**, or correlated (Gardner *et al*, 2000), not conditionally independent. (**Note** If either the  $Se$  or  $Sp$  of a test equals 1 (*ie* is perfect), then it will always be conditionally independent (for that characteristic) of other tests.)

The extent of the dependence can be calculated as shown below and in Example 5.9.

1. Denote the observed proportion of  $D+$  animals with a positive test result to both tests as  $p_{111}$  (more generally  $p_{ijk}$ ;  $i$  denoting test 1 result,  $j$  denoting test 2 result, and  $k$  denoting disease status (1=diseased, 0=non-diseased)).
2. In the  $D+$  group, and using the sample estimates of  $Se$  for tests 1 and 2 respectively, ( $Se_1$  and  $Se_2$ ), the covariance is:

$$\text{covar}(+) = p_{111} - Se_1 * Se_2 \tag{Eq 5.20}$$

3. Similarly, in the  $D-$  group and using the sample estimates of  $Sp_1$  and  $Sp_2$ , the covariance is:

$$\text{covar}(-) = p_{000} - Sp_1 * Sp_2 \tag{Eq 5.21}$$

The usual circumstance would be that these covariances would be positive, indicating dependence. In a more formal sense, if one calculates an odds ratio ( $OR$ ) on the data from the  $D+$  group ( $OR+$ ) and separately on the  $D-$  group ( $OR-$ ), these  $OR$ s describe the above 2 covariances respectively, because, if the tests were conditionally independent, the  $OR$ s would equal 1. Similarly, if the test results are conditionally independent, the kappa statistic in data from  $D+$  and  $D-$  individuals would both equal 0.

4. Given dependence, the  $Se$  and  $Sp$  resulting from parallel interpretation of 2 tests are:

$$Se_p = 1 - p_{001} = 1 - (1 - Se_1) * (1 - Se_2) - \text{covar}(+) \tag{Eq 5.22}$$

$$Sp_p = p_{000} = Sp_1 * Sp_2 + \text{covar}(-) \tag{Eq 5.23}$$

From series interpretation of 2 tests these are:

$$Se_s = p_{111} = Se_1 * Se_2 + \text{covar}(+) \tag{Eq 5.24}$$

$$Sp_s = 1 - p_{110} = 1 - (1 - Sp_1) * (1 - Sp_2) - \text{covar}(-) \tag{Eq 5.25}$$

Functionally, this means that the gains/losses from using either of these approaches are not as great as predicted under conditional independence. It can also affect the choice of tests to be used. For example, a more optimal outcome might arise from combining 2 independent tests with lower sensitivities than 2 dependent tests with higher sensitivities.

**Example 5.9 Estimating covariance between test results**

data = ISA\_test

Using the  $Se$  and  $Sp$  estimates obtained in Example 5.8, the covariance in the  $D+$  and  $D-$  groups are:

$D+$  group:  $\text{covar}(+) = p_{111} - Se_1 * Se_2 = 0.761 - 0.726 = 0.035$

$D-$  group:  $\text{covar}(-) = p_{000} - Sp_1 * Sp_2 = 0.930 - 0.931 = -0.001$

There is a slight positive covariance in the  $D+$  group, but it is sufficiently small that the correction will not materially affect the results when the tests are used in combination. There is virtually no covariance in the  $D-$  group.

## 5.7 EVALUATION OF DIAGNOSTIC TESTS

There is a variety of approaches to estimating the  $Se$  and  $Sp$  of a diagnostic test. These include:

- the use of gold standard populations
- the use of a gold standard reference test
- the use of a pseudo-gold standard test (or combination of tests)
- reference test with known  $Se$  and  $Sp$
- evaluation when there is no 'gold standard' (Section 5.8).

### 5.7.1 Gold standard populations

In some situations, a population which is assumed to be completely free of a disease may be available for estimating the  $Sp$  of a test. For example, one of 4 populations sampled in an evaluation of diagnostic tests for ISA was assumed to be free and provided a direct estimate of  $Sp$  (Nerette *et al*, 2005a). The main issue to be considered in this case is whether or not the characteristics of the population result in an estimate of  $Sp$  that is appropriate for the population of interest. It is not often possible to identify a population in which all animals are assumed to be  $D+$  for the estimation of  $Se$ .

Another approach to estimating  $Sp$  when disease is known to be infrequent (say, less than 2%) is to assume that all of the test positive animals are false positives (*ie*  $Sp=1-AP$ ). For example, in Ireland, about 4 animals per 1,000 test positive to the skin test for bovine tuberculosis; hence, the  $Sp$  of this test cannot be less than  $1-0.004=0.996$  (99.6%). If a portion of the test positives are found (or known) to be true positives, then the  $AP$  can be adjusted accordingly and the estimate of the  $Sp$  raised.

### 5.7.2 Gold standard reference test

In some cases, a gold standard test (or combination of tests) is available. Studies using a gold standard reference test may be conducted in one of 2 ways. One approach (**1-stage approach**) is to test a sample of animals from the population with both the gold standard test(s) and the test being evaluated.  $Se$  and  $Sp$  can then be computed directly and the binomial distribution can be used to calculate the standard errors and confidence limits (see Section 4.10). A drawback of this approach is that a very large sample size will be required to obtain a reasonable estimate of  $Se$  if the disease prevalence is low.

An alternative **2-stage approach** is to screen a sample from the population with the test being evaluated and then a subsample of  $T+$  and  $T-$  animals is submitted to the gold standard test (to determine their 'true' health status). It is vitally important that selection of animals for verification be independent of their true health status (random sampling is the preferred method). If the fraction of  $T+$  animals that is selected for verification is different than that fraction of  $T-$  samples, this must be taken into account when estimating  $Se$  and  $Sp$ . If we denote the fraction ( $sf$ ) of the test positives that are verified as  $sf_{T+}$ , and that of the test negatives as  $sf_{T-}$ , then the corrected estimate of  $Se$  is:

$$Se_{\text{corr}} = \frac{a/sf_{T+}}{a/sf_{T+} + b/sf_{T-}} \quad \text{Eq 5.26}$$

and the corrected estimate of  $Sp$  is:



$$Sp_{\text{corr}} = \frac{d / sf_{T-}}{d / sf_{T-} + c / sf_{T+}} \tag{Eq 5.27}$$

(See Example 5.10.) If  $sf_{T+}=sf_{T-}$ , no adjustment for the sampling fractions is needed.

The variances of these ‘corrected’ proportions are calculated using only the number of verified individuals in the variance formulae (*ie* the  $a+b$  verified animals for  $Se_{\text{corr}}$  and the  $c+d$  verified animals for  $Sp_{\text{corr}}$  (Table 5.2)) (Greiner & Gardner, 2000a).

Procedures for deciding the optimal balance between animals tested with the new test (stage 1) and animals submitted to gold standard testing (stage 2) have been published (McNamee, 2002). A procedure in which stage 2 is replaced with a sequential process of first evaluating the specificity of the test and then (if the specificity is acceptable) proceeding to evaluating the  $Se$  has been recommended (Wruck *et al*, 2006).

Regardless of whether a one-stage or 2-stage approach is used, it is advantageous to have a spectrum of host attributes and clustering units (if any) present (*ie* animals from a number of different farms). The results should be assessed for differences in  $Se$  or  $Sp$  by host attributes using logistic regression (see Section 5.9.2). Blind assessment and complete work-ups of all animals are useful aids to prevent bias in the estimates. When  $Se$  and  $Sp$  are estimated based on samples obtained from several animals within a number of farms, adjustment of the SEs for the clustering effect should be made. This can be done using hierarchical multilevel procedures (Chapter 20 and 22) or survey statistics (Chapter 2) (Greiner, 2003).

**Example 5.10 Estimating  $Se$  and  $Sp$  using a validation subsample**

data = none

Suppose that, at slaughter, we examine 10,000 cattle for the presence of lesions consistent with bovine tuberculosis (TB). We find lesions in 242 cattle. A detailed follow-up is done on 100 of the animal specimens with lesions and on similar tissue specimens from 200 of the ‘clear’ animals. In the animals with lesions, 83 are confirmed as bovine tuberculosis, whereas 2 of the 200 clear animals are found to have tuberculosis. The data are shown here.

		Lesion+	Lesion-
TB+	(D+)	83	2
TB-	(D-)	17	198
		100	200

and

$$sf_{T+} = 100/242 = 0.413$$

$$sf_{T-} = 200/9758 = 0.020$$

From these we can calculate  $Se_{\text{corr}}$  and  $Sp_{\text{corr}}$

$$Se_{\text{corr}} = \frac{83/0.413}{83/0.413 + 2/0.0205} = \frac{200.9}{298.5} = 0.673$$

with approximate SE of  $\sqrt{(0.673*0.328)/85} = 0.051$  and

$$Sp_{\text{corr}} = \frac{198/0.0205}{198/0.0205 + 17/0.413} = \frac{9658.5}{9941.2} = 0.996$$

with approximate SE of  $\sqrt{(0.996*(1-.996))/215} = 0.004$

### 5.7.3 Pseudo-gold standard procedures

Pseudo-gold standards involve the use of a combination of imperfect tests as a substitute for a gold standard. Two approaches have been described: **discrepant resolution** and **composite reference standard**. The former has a problem in that disease status measurement is conditional upon the test being evaluated and hence, produces biased results (Miller, 1998). It will not be considered further.

A composite reference standard (CRS) is formed by first testing all samples with a reference test and then all reference test negative samples are tested with a resolver test. The results are interpreted in parallel so that any specimen that was positive on either the reference or resolver test is considered CRS positive while specimens that are negative on both tests are CRS negative (Alonzo & Pepe, 1999). These results are then used to evaluate the test of interest in place of a gold standard test.

Pseudo-gold standards can also be created using an ad-hoc, study-specific approach provided there is sufficient justification for the approach chosen (Nerette *et al*, 2008). Example 5.11 shows the use of a composite reference standard for evaluating the  $Se$  and  $Sp$  of a test.

### 5.7.4 Reference test with known $Se$ and $Sp$

If the  $Se$  and  $Sp$  of a reference test ( $Se_{ref}$  and  $Sp_{ref}$ , respectively) are known, then from the data in a 2X2 table based on the new test results (but with disease status determined by the reference test), we could estimate the  $Se_{new}$  and  $Sp_{new}$  of the new test using the syntax of Table 5.2 as follows (Enøe *et al*, 2000):

$$Se_{new} = \frac{n_1 Sp_{ref} - c}{n Sp_{ref} - m_0} \quad Eq\ 5.28$$

$$Sp_{new} = \frac{n_0 Se_{ref} - b}{n Se_{ref} - m_1} \quad Eq\ 5.29$$

We could also estimate  $P$  using

$$P = \frac{n(Sp_{ref} - 1) + m_1}{n(Se_{ref} + Sp_{ref} - 1)} \quad Eq\ 5.30$$

Variance formulae are available (Gart & Buck, 1966). This procedure assumes that, conditional on the true disease state, the new test and the reference test are independent, which may not be a valid assumption.

## 5.8 EVALUATION WHEN THERE IS NO GOLD STANDARD

In situations in which there is neither a reasonable gold standard, nor a test(s) with known characteristics ( $Se$  and  $Sp$ ), latent class models can be used to simultaneously estimate the  $Se$  and  $Sp$  of 2 or more tests without any assumption about the true disease status of each individual (Hui & Walter, 1980). There has been a large body of literature published in recent years on the use of latent class models for evaluating diagnostic tests. This section will provide only a brief introduction along with some selected references for further reading.

**Example 5.11 Use of pseudo-gold standard for evaluating *Se* and *Sp* of a diagnostic test**  
 data = isa\_lcm

The data used in this example and Examples 5.12 and 5.13 are a subset of those derived from an evaluation of diagnostic tests for infectious salmon anemia (ISA)(Nerette *et al*, 2005a). Samples were collected from 4 populations of fish: high prevalence = moribund fish in cages experiencing an outbreak, medium prevalence = healthy fish from outbreak cages, low prevalence = healthy fish from cages near an outbreak (but no ISA evident in the cage) and zero prevalence = a population assumed to be completely free of disease. Four tests were used on each fish: an indirect fluorescent antibody test (IFAT), 3 polymerase chain reaction (PCR) tests (in 3 different laboratories) (PCRa, PCRb and PCRc) and 2 virus isolation (VI) tests (VIa and VIb).

A composite reference standard (CRS) test result was computed for each fish using PCRb as the reference and VI as the resolver test. The data for this calculation are shown below. This was used to estimate the *Se* and *Sp* of the IFAT.

	Reference test		Resolver test		CRS	
	PCRb		VIa		CRS	
IFAT	1	0	1	0	1	0
1	56	10	6	4	62 (=56+6)	4
0	29	234	13	221	42 (=29+13)	221
		244			104	225

The 244 samples that were PCRb- were evaluated using the resolver test (VI); 13 of the 234 PCRb-, IFAT- samples were positive on VI and were added to the CRS+ group; 6 of the 10 PCRb-, IFAT+ samples were positive and were also added to the CRS+ group.

The *Se* of IFAT was estimated to be  $62/104 = 0.60$ , while the *Sp* was  $221/225 = 0.98$ .

**5.8.1 Latent class models—principles and assumptions**

**Latent class models** (LCM) involve an unknown (latent) variable that takes categorical values. In this case, the unobserved variable is the true disease status of each animal which is usually assumed to be binary (diseased or non-diseased). Such models can be used to evaluate the accuracy of diagnostic tests when there is no gold standard. In its standard and most commonly used form, the model involves 3 assumptions: (i) the target population should consist of 2 (or more) subpopulations with different prevalences; (ii) the sensitivity (*Se*) and specificity (*Sp*) of the diagnostic tests should be constant across subpopulations (*ie* the ability of a test to detect infected individuals should be the same regardless of whether the test is used in a population with a high prevalence of infection or one with a low prevalence); and (iii) the tests should be conditionally independent given the disease status (Enøe *et al*, 2000). (For a discussion of conditional independence, see Section 5.6.2)

If the data consisted of test results from 2 tests applied to individuals from 2 populations, they can be presented as shown in Table 5.4.

**Table 5.4 Layout of data for evaluating Se and Sp using latent class models (2 populations and 2 tests)**

	Population 1				Population 2		
	T <sub>2</sub> <sup>+</sup>	T <sub>2</sub> <sup>-</sup>	Total		T <sub>2</sub> <sup>+</sup>	T <sub>2</sub> <sup>-</sup>	Total
T <sub>1</sub> <sup>+</sup>	$n_{kij} = n_{111}$	$n_{112}$	$n_{11.}$	T <sub>1</sub> <sup>+</sup>	$n_{211}$	$n_{212}$	$n_{21.}$
T <sub>1</sub> <sup>-</sup>	$n_{121}$	$n_{122}$	$n_{12.}$	T <sub>1</sub> <sup>-</sup>	$n_{221}$	$n_{222}$	$n_{22.}$
Total	$n_{1.1}$	$n_{1.2}$	$n_1$	Total	$n_{2.1}$	$n_{2.2}$	$n_2$

The distribution of tests results are distributed according to a multinomial model for the observed counts in each population:

$$(n_{kij}) \sim \text{multinomial} (n_k, p_{kij})$$

where  $n_k$  is the sample size in population  $k$  ( $k=1,2$ ) and  $p_{kij}$  is the probability of an animal being in this cell ( $i$  and  $j$  represent the 2 tests;  $i, j = 1, 2 \sim +, -$ ).

If  $\theta_k$  is the true (unknown) prevalence in population  $k$ , then assumptions (ii) and (iii) lead to:

$$\begin{aligned}
 p_{111} &= \theta_1 Se_1 Se_2 + (1 - \theta_1)(1 - Sp_1)(1 - Sp_2) \\
 p_{112} &= \theta_1 Se_1(1 - Se_2) + (1 - \theta_1)(1 - Sp_1)Sp_2 \\
 p_{121} &= \theta_1(1 - Se_1)Se_2 + (1 - \theta_1)Sp_1(1 - Sp_2) \\
 p_{111} &= \theta_1(1 - Se_1)(1 - Se_2) + (1 - \theta_1)Sp_1Sp_2
 \end{aligned}
 \tag{Eq 5.31}$$

Consequently, the LCM contains 6 parameters: the  $Se$  and  $Sp$  of each test and the prevalence ( $\theta$ ) in each population. Given that the population sample sizes are fixed (by the study design), these 2 tables have a total of 6 df (each table contributes 3 df because once the value of 3 cells is known, the fourth can be computed by subtraction). Consequently, in this particular situation, estimation involves reparameterising the 6 observed values into 6 parameter estimates and there are no residual df which can be used to evaluate the model fit and validity. With more than 2 tests and/or 2 populations, the LCM involves a reduction in parameters relative to the full multinomial model and the residual df can be used to assess the fit of the model.

### 5.8.2 Estimation procedures

Both maximum likelihood (ML) and Bayesian estimation procedures can be used to fit LCMs (see Enøe *et al* (2000); Hui & Zhou (1998) for reviews of the earlier literature). ML estimates are a set of parameter estimates that were most likely to have generated the observed data and are obtained by maximising the likelihood function. Software for obtaining ML estimates using the ‘TAGS’ program (Pouillot *et al*, 2002), along with software for many other approaches discussed in this section can be obtained through: <http://www.epi.ucdavis.edu/diagnostictests>. Example 5.12 shows the results from the use of a latent class model to estimate the  $Se$  and  $Sp$  of 3 tests for ISA in salmon.

ML estimation is usually carried out using an Expectation-Maximisation (EM) algorithm which is a general estimation procedure for problems involving incomplete data (in this situation it is the latent variable which is missing). Following this, a Newton-Raphson estimation step is carried out to produce estimates of the SEs. There are several methods for obtaining confidence

**Example 5.12 Evaluation of *Se* and *Sp* using a latent class model**

data = isa\_lcm

A subset of the data from the ISA test evaluation (see Example 5.10) were used to evaluate the test characteristics of the IFAT (dichotomised), PCRa and VIa tests. Data from 3 populations (high, medium and low prevalence populations) were used and maximum likelihood estimates of the *Se* and *Sp* of each test and P of ISA in each population obtained. The data were as follows:

IFAT	PCR1	VI	Number of fish (population)		
			High P	Medium P	Low P
0	0	0	6	49	77
1	0	0	0	1	2
0	1	0	13	21	19
1	1	0	1	0	0
0	0	1	0	0	0
1	0	1	0	0	0
0	1	1	22	9	1
1	1	1	57	20	1

ML estimates (and 95% CI) of the parameters were obtained using the TAGS software.

	Prevalence			IFAT		PCR1		VI	
	Low	Med	High	Se	Sp	Se	Sp	Se	Sp
<b>Estimate</b>	0.020	0.292	0.820	0.702	0.984	1.000	0.724	0.979	1.000
<b>Lower CI</b>	0.005	0.211	0.720	0.608	0.950	na	0.653	0.842	na
<b>Upper CI</b>	0.077	0.389	0.890	0.782	0.995	na	0.785	0.998	na

The most surprising result was the low *Sp* of the PCR but additional evaluations showed that this was in fact the case. It was not possible for the TAGS software to provide a plausible CI in situations where the point estimate was 1.000. Because the number of df available from the data ( $3 \times (2^3 - 1) = 21$ ) was greater than the number of parameters estimated (9), there were residual df that could be used to evaluate how well this LCM fit the data. The deviance (19.5) on 12 df had a P value of 0.08 providing some marginal evidence that there were problems with the fit of the model. More detailed evaluation of this problem is beyond the scope of this text (see Nerette *et al* (2008)).

intervals, but the most commonly used is bootstrapping. ML estimation does not perform well in small sample situations, particularly if there are many data cells with small or zero frequencies (Walter, 2005a).

Alternatively, the Bayesian approach can be used to combine *a priori* scientific knowledge about unknown parameters with the information contained in the likelihood based on the observed data. (Bayesian methods are discussed in general in Chapter 24.) A detailed discussion about Bayesian procedures for fitting LCMs is beyond the scope of this text. However, the advantages offered by Bayesian procedures are as follows.

- Bayesian models are very flexible and it is relatively easy to extend the models to account for factors such as dependence among test results.
- If prior information about any of the parameters (*Se*, *Sp* or prevalences) is available, it

can be incorporated into the analysis. This effectively increases the df available and facilitates:

- model-building when the observed data are inadequate to provide good estimates of the required parameters, and
- evaluating the model (*eg* obtaining goodness-of-fit estimates when they would not otherwise be possible by providing prior information about some of the parameters).

Overviews of Bayesian estimation (Branscum *et al*, 2005; Joseph *et al*, 1995) have been published and some recent veterinary examples are Engel *et al* (2008); Georgiadis *et al*, (2003); Kostoulas *et al* (2006).

Although not always possible, it is important to evaluate the assumptions underlying the LCM. Approaches for assessing the overall fit of the model and for dealing with each of the 3 assumptions are considered here.

### 5.8.3 Goodness-of-fit

If the number of degrees of freedom in the data exceeds the number in the LCM, it is possible to obtain an estimate of the goodness-of-fit of the model.

Pearson residuals can be computed for each cell in the data tables by comparing the predicted value from the LCM ( $n_k p_{kij}$ , where  $p_{kij}$  is the estimated cell probability) to the observed value ( $n_k$ ). These may then be normalised by division by the estimated SE of the predicted value to yield.

$$\epsilon_{kij} = (n_{kij} - n_k p_{kij}) / \sqrt{n_k p_{kij}} \quad \text{Eq 5.32}$$

The squared sum of these residuals is assumed to follow a  $\chi^2$  distribution although the exact reference distribution is unknown. Although this test is assumed to have relatively little power for detecting lack of fit, numerically large individual residuals identify cells with an apparent lack of fit.

Alternatively, the deviance can be computed as twice the difference between the log likelihood of the full multinomial model and the latent class model. The deviance can be compared with a  $\chi^2$  distribution (see Chapter 16 for a discussion of likelihood ratio tests). Our experience is that this test often produces evidence of a statistically significant lack of fit even when the estimates appear reasonable (see Example 5.12).

### 5.8.4 Prevalence of the 2 populations differs

The greater the difference in the prevalences among the populations studied, the more precise the estimates of *Se* and *Sp* will be. Consequently, it is desirable to identify populations in which radically different prevalences are expected. In the ISA study of Examples 5.11–5.13 this problem was approached by sampling fish from populations ranging from moribund fish collected during a disease outbreak to completely healthy fish from cages with no evidence of clinical disease. In cases where samples are only available from a single population, it may be possible to stratify the population on the basis of some characteristic which is expected to influence disease prevalence. For example, Nielsen *et al* (2002) used herd size, geographic location (zip code) and herd veterinarian as criteria for dividing a single population of dairy cattle into groups with assumed differences in prevalence. However, care must be taken to

ensure that the prevalences are truly different.

### 5.8.5 *Se* and *Sp* constant across populations

If the *Se* and/or *Sp* of a test varies across populations in a study, the overall estimate provided by the LCM will be some mixture of population specific estimates and will be weighted toward the population which provides the most information about the parameter. For example, if the *Se* of a test varies across populations, the overall estimate of the *Se* will be weighted toward the estimate in the high prevalence population because it contains the most *D+* animals and hence, provides the most information about the *Se* of the test (Toft *et al*, 2005).

If a pseudo-gold standard test result can be obtained, then the data can be divided into *D+* and *D-* datasets and regression procedures (described in Section 5.9.2) can be fit with individual test results as the outcome. Inclusion of population identifiers in the model will provide evidence as to whether or not the population affects the estimates of *Se* and *Sp*.

Alternatively, Bayesian procedures with informative priors can be used to fit LCMs individually for each population. Informative priors are required because, assuming 2 tests are being evaluated, a single population only provides 3 df but 5 are required for the LCM. Consequently, at least 2 informative priors will have to be included in the model.

### 5.8.6 Dependence among test results

Tests are more likely to be independent if they have very different biological bases (*eg* a culture procedure such as virus isolation and a molecular technique such as PCR). However, this is not necessarily sufficient to guarantee independence.

Once again, if 3 or more tests have been applied, pseudo-gold standards can be used to evaluate dependence among test results by dividing the data into *D+* and *D-* individuals based on the pseudo-gold standard. Log-linear models can be used to compare nested models to determine the most parsimonious dependence structure with minimal, non-significant loss of fit for the data (Hanson *et al*, 2000).

Latent class models can be extended to account for dependence among tests in order to relax the assumption of conditional independence (*eg* Branscum *et al*, 2005; Dendukuri & Joseph, 2001; Georgiadis *et al*, 2003). If more complex models, which allow for dependence between tests, fit the observed data better, then it is assumed that the tests are not independent and the estimates from the more complex model are preferred. Albert and Dodd (2004) showed that in many practical situations, ML estimators of *Se* and *Sp* are biased when the dependence structure is misspecified, and that it is difficult to choose the correct dependence structure using likelihood comparisons and other model diagnostics. They demonstrated that several models may fit the data equally well, while providing different accuracy estimates. They recommended using a gold standard whenever possible or even collecting gold standard information on a fraction of subjects to aid in choosing a model.

## 5.9 OTHER CONSIDERATIONS IN TEST EVALUATION

### 5.9.1 Factors that affect $Se$ and $Sp$

Sensitivity and specificity represent average values of the test characteristics and as such, we can expect their levels to vary from one subgroup of the population to another. Consequently, when estimating  $Se$  and  $Sp$ , it is important that the study sample to which the gold standard diagnostic procedure is applied be representative of the target population (*ie* those animals to whom the test will be applied in the future). This representativeness refers to the attributes of the animals being tested including their age, breed, sex *etc*, as host and environmental factors might influence the ability of a test to detect disease. In fact, often it is useful to stratify the results based on the more important of these factors in order to obtain more valid stratum-specific estimates. In addition, it is important that the study group contains an appropriate spectrum of disease (*eg* severity, chronicity or stage of development). Certainly, the test characteristics might differ in different stages of the disease process; for example, tests for Johne's disease work much better once the animal is clinically ill as opposed to only being infected with the organism *Mycobacterium avium* subsp *paratuberculosis* (Map).

While the  $Se$  and  $Sp$  are often considered characteristics of a test, there is increasing evidence that for many tests, the  $Se$  and  $Sp$  vary with the characteristics of the population to which they are applied (Greiner & Gardner, 2000a). For example, the specificity of serologic tests for *Brucella abortus* is higher when the test is used in populations in which no calftlood vaccination is used compared with vaccinated populations (Dohoo *et al*, 1986). Often it is important to know what characteristics of a population affect the  $Se$  and  $Sp$  of a test (some might prefer to think of factors relating to the occurrence of false negative or false positive results).

### 5.9.2 Evaluating effects of factors on $Se$ and $Sp$

If there are few factors that affect  $Se$  and  $Sp$ , you can stratify on these and estimate the  $Se$  and  $Sp$  in each stratum. However, when there are several factors to investigate, stratification rapidly runs into problems of inadequate sample size and it is more convenient to use a logistic regression approach (Coughlin *et al*, 1992). For details on logistic regression see Chapter 16.

The logistic regression approach involves modelling the dichotomous test outcome (positive or negative) as a function of the true disease status variable ( $X_{is}$ ) as well as the factors that might affect the  $Se$  and  $Sp$ . This can either be done by carrying out separate logistic regressions using the  $D+$  and  $D-$  animals (as shown in the equations below and in Example 5.13) or by including the true disease status variable ( $X_{is}$ ) in the model. In the latter approach, it might be necessary to include interaction terms between  $X_{is}$  and the other factors to allow for the fact that those factors might have different effects in  $D+$  and  $D-$  animals. Non-significant factors might be eliminated, but the variable representing the true disease status of the animal must remain in the model.

For a given set of factor values, the  $Se$  of the test will be:

$$Se = \frac{e^{\mu^+}}{1 + e^{\mu^+}} \quad \text{Eq 5.33}$$

where  $\mu^+ = \beta_0 + \sum \beta_j X_j$  is the linear predictor from a logistic model based only on  $D+$  animals.



The specificity of the test is:

$$Sp = 1 - \frac{e^{\mu^-}}{1 + e^{\mu^-}} \quad \text{Eq 5.34}$$

where  $\mu^- = \beta_0 + \sum \beta_j X_j$  is the linear predictor from a logistic model based only on *D*- animals.

One can use a similar approach to estimate predictive values but in that case, the outcome is the true disease status and the test result is one of the explanatory variables. Examples of this are discussed elsewhere (Greiner & Gardner, 2000a). Example 5.13 shows the use of logistic regression to evaluate the effect of population on estimates of *Se* and *Sp* in the ISA data.

### 5.9.3 Clustering of test results

In addition to considering how population characteristics may influence estimates of *Se* and *Sp*, it is important to take into consideration the fact that observations used in validation studies may be clustered (observations not independent). For example, data may come from cows which are housed in a number of herds. Procedures for dealing with clustered data are described in more detail in Chapters 20-23. One approach to dealing with the lack of independence would be to include random effects for clustering variables (*eg* random herd effects) in the regression modelling approaches described above.

## 5.10 SAMPLE SIZE REQUIREMENTS

### 5.10.1 Gold-standard based procedures

When designing a study to estimate the *Se* and/or *Sp* of a test, we need to consider the number of animals that is required to obtain a specified precision for each estimate. These form the basis for estimating the 95% (or other specified level) CIs as shown in Example 5.4. For *Se*, estimates within  $\pm 5\%$  might suffice, whereas for screening low-risk populations, much larger sample sizes are needed as *Sp* estimates need to be within at least  $\pm 0.5\%$  of the true value. In a diagnostic setting, *Sp* estimates within 3-5% of the true value should suffice. As these estimates of *Se* and *Sp* are binomial proportions, sample size formulae for estimating a binomial proportion (see Chapter 2) are applicable.

### 5.10.2 Latent class models

In general, sample size requirements for studies using LCM to estimate *Se* and *Sp* are much larger than for those based on a gold standard approach. A spreadsheet for the calculation of sample sizes in the situation of 2 conditionally independent tests applied to 2 populations is available (Georgiadis *et al*, 2005). It confirms that sample size is heavily influenced by the magnitude of the difference between the prevalences of disease in the 2 populations.

## 5.11 HERD-LEVEL TESTING

If a herd, or other aggregate of individuals, is the unit of concern, and a single test of the herd (*eg* a culture of a bulk-tank milk sample for *Strep. agalactia* in a dairy herd) is taken to classify

**Example 5.13 Evaluation of factors affecting *Se* and *Sp***

data = isa\_lcm

Separate logistic regression models were fit to the CRS+ and CRS- fish with population as the only predictor. The model for the CRS+ fish (which estimates sensitivity) is shown below. (**Note** The population free of disease was dropped because it provided no information about the *Se* of the test).

Logistic regression Number of obs = 104  
LR chi2(2) = 3.07  
Prob > chi2 = 0.2155  
Pseudo R2 = 0.0 219  
Log likelihood = -68.6

IFAT	Coef	SE	Z	P> z	95% CI	
prev=med	0.598	1.439	0.420	0.678	-2.222	3.418
prev=high	1.743	1.472	1.180	0.236	-1.143	4.628
Constant	-3.689	1.012	-3.640	0.000	-5.673	-1.705

The likelihood ratio test for the overall significance of population as a predictor was 3.07 with 2 df ( $P=0.22$ ), indicating that population did not have a significant effect on the *Se* of the test. A similar model (not shown) of CRS- fish confirmed that the population did not have a significant effect on *Sp* either ( $P=0.64$ ).

The models were extended to account for clustering of test results within cages by the addition of random effects for cages. Not all fish had their cage identifier recorded which limited the sample size available for this analysis. There was marginal evidence of clustering of positive test results, but none for negative test results (results not shown).

the herd as test positive or test negative, the previously described approach to test evaluation and interpretation applies directly. The herd becomes the unit of concern rather than the individual. (**Note** Throughout this section, the term ‘herd’ will be used but the reader should recognise that this could be any identifiable group of individuals.)

However, frequently we are asked to certify the health status of a herd based on test results compiled from a number of individuals. In this instance, in addition to the *Se* and *Sp* of the test at the individual level, 3 factors interplay in determining the *Se* and *Sp* at the herd level (denoted *HSe* and *HSp*)—namely, the frequency of disease within infected herds, the number of animals tested in the herd, and the number of reactor animals per herd that will designate a positive or negative herd. Once the *HSe* and *HSp* of the procedure are known, the evaluation of the predictive values of positive and negative herd results follows the same pattern as already described (Christensen & Gardner, 2000; Martin *et al*, 1992).

### 5.11.1 Apparent prevalence

As mentioned, herd sensitivity (*HSe*) and herd specificity (*HSp*) are influenced by the individual level *Se* and *Sp*, the within herd *P*, and the threshold number, or percentage, of positive tests that denote the herd, as test positive. For simplicity, we assume only one test is used; however, multiple tests and repeat testing results can make up the herd test and one need only establish their combined *Se* and *Sp*. Within a herd, the probability of obtaining a positive test is:

$$AP = p(T+) = P * Se + (1 - P)(1 - Sp) \quad \text{Eq 5.35}$$

If a herd is infected, then one or more positive test results may arise correctly based on  $P*Se$ , or may arise correctly, but for incorrect reasons, because of the  $(1-P)(1-Sp)$  component.

Thus, if disease is present, the  $AP$  is:  $AP_{pos}=P*Se+(1-P)(1-Sp)$

However, if the herd is not infected ( $P=0$ ) then the  $AP$  is:  $AP_{neg}=(1-Sp)$ .

**5.11.2 Herd sensitivity**

If the critical number of animals testing positive to denote the herd as test positive is  $k$ , we can use a suitable probability distribution for  $AP$  and solve for the probability of  $\geq k$  animals testing positive when  $n$  animals are tested. If  $n/N$  is less than 0.1, then a binomial distribution is acceptable for sampling of  $n$  animals from a total of  $N$  animals in a herd; otherwise, the hypergeometric distribution, which provides more accurate estimates, should be used. In the simplest setting, if  $k=1$ , the easiest approach is to compute the binomial probability for  $k=0$  and take 1 minus this probability to obtain the probability of one or more test positive animals. Thus for  $k=1$  and assuming the herd is infected:

$$HSe = 1 - (1 - AP_{pos})^n \tag{Eq 5.36}$$

In the more general case, if  $k$  or more positives are required before a herd is declared positive, the  $HSe$  can be estimated as:

$$HSe = 1 - \sum_0^{k-1} C_{k-1}^n (AP_{pos})^{k-1} (1 - AP_{pos})^{n-(k-1)} \tag{Eq 5.37}$$

where  $C_k^n$  is the number of combinations of  $k$  positives out of  $n$  animals tested.

**5.11.3 Herd specificity**

If the herd is disease free and  $k=1$ , then

$$HSp = Sp^n$$

More generally, at a cutpoint of  $k$  or more positives, the  $HSp$  will be:

$$HSp = \sum_0^{k-1} C_{k-1}^n (Sp)^{n-(k-1)} (1 - Sp)^{k-1} \tag{Eq 5.38}$$

Both  $HSe$  and  $HSp$  are estimates of population parameters that apply to herds with the underlying conditions and characteristics used to determine the estimates.

**5.11.4 Relationships among  $Se$ ,  $Sp$ ,  $HSe$  and  $HSp$**

Some general findings from studying herd test characteristics are:

1. If  $n$  is fixed,  $HSe$  increases with  $P$  and/or  $AP$ , provided  $Se > (1-Sp)$ .
2. As  $n$  increases,  $HSe$  increases. Gains in  $HSe$  from increasing  $n$  are especially large if  $AP < 0.3$ .
3. With fixed  $n$ ,  $HSe$  increases as  $Sp$  decreases (noted earlier).
4.  $HSp$  decreases as  $Sp$  decreases or as  $n$  increases.

A program called Herdacc (©D Jordan, 1995) is available at [www.vetschools.co.uk/EpiVetNet/software.htm](http://www.vetschools.co.uk/EpiVetNet/software.htm) to perform 'what-if' calculations to see how changing the sample size, the number required to consider a herd positive or the statistical distribution (binomial or hypergeometric) affects the results. An example of estimating  $HSe$  and  $HSp$  is shown in Example 5.14.

### 5.11.5 Uncertainty in estimates of $Se$ , $Sp$ and prevalence

It is rare that the  $Se$  and  $Sp$  of the test(s) being used, or the underlying individual animal prevalence, are known with certainty. Consequently, there will be uncertainty in the estimate of  $HSe$  and  $HSp$ . One approach to accounting for uncertainty in  $Se$  and  $Sp$  is to compute the variance of the estimate of the  $AP$  using the following formula (Rogan & Gladen, 1978):

$$var(AP) = P^2 * \frac{Se*(1-Se)}{N} + (1-P)^2 * \frac{Sp*(1-Sp)}{M} \quad \text{Eq 5.39}$$

#### Example 5.14 Estimating $HSe$ and $HSp$

We will assume that we are testing herds with an average of 60 adult cattle for the presence of *Mycobacterium avium* subsp *paratuberculosis* (Map) using an ELISA with an estimated  $Se$  of 0.391 and  $Sp$  of 0.964. We will assume that if Map is present, the true prevalence at the time of testing is 12%. Thus the  $AP$  in the herds with disease will be:

$$AP_{pos} = p(T+) = P * Se + (1-P)(1-Sp) = 0.12 * 0.391 + (0.88)(1-0.964) = 0.079$$

and the  $AP$  in the disease-free herds will be:  $AP_{neg} = 0.036$

Now, assume that the critical number of positive-testing animals to denote a herd as test positive is  $Y \geq 2$ . For the purposes of this example, we will use the binomial probability distribution to solve for the probability of  $\$2$  positive-testing animals when  $n=60$  animals are tested (assuming an infinite population). The probability of  $Y \geq 2$  is found by first computing the probability that  $Y < 2$ .

$$p(Y < 2) = \sum_0^{Y-1} C_Y^n AP^Y (1-AP)^{n-Y}$$

$$\text{The probability that } Y = 0 \text{ is: } p(Y=0) = C_0^{60} * (0.079)^0 * (1-0.079)^{60} = 0.007$$

$$\text{The probability that } Y = 1 \text{ is: } p(Y=1) = C_1^{60} * (0.079)^1 * (1-0.079)^{59} = 0.037$$

The sum of these 2 probabilities is 0.044. Hence, the probability of 2 or more animals testing positive in a herd with  $P = 0.12$  is  $1-0.044 = 0.956$ , which gives us the  $HSe$  estimate.

For  $HSp$ , we would assume the herds are disease free so:

$$\text{the probability that } Y = 0 \text{ is: } p(Y=0) = C_0^{60} * (0.964)^{60} * (1-0.964)^0 = 0.111$$

$$\text{the probability that } Y = 1 \text{ is: } p(Y=1) = C_1^{60} * (0.964)^{59} * (1-0.964)^1 = 0.248$$

Hence the  $HSp$  is  $0.111 + 0.248 = 0.359$ .

With an  $HSe$  of 96%, we can be confident that we will declare the herd as infected if it is infected. However, with the  $HSp$  of only 36%, we will declare 64% of Map-free herds as infected, so the test would need to be used with great care.

where  $N$  and  $M$  are the number of true positive and true negative animals, respectively. A confidence interval of the estimate of  $AP$  can then be computed and the lower and upper limits used in the formula for  $HSe$  (Eq 5.37) to obtain a confidence interval for  $HSe$ .

Similarly a confidence interval for  $HSp$  can be built using the lower and upper limits of the confidence interval for  $Sp$  (see Section 4.10).

The approach described above does not take into account that disease is likely to cluster within herds and given different disease processes within herds, the  $Se$  and  $Sp$  of the test may also vary from herd to herd. A Monte Carlo simulation program for evaluating herd-level test performance taking these factors into consideration has been published (Jordan & McEwen, 1998) and recently used to estimate herd-level test characteristics for tuberculosis testing programs (Norby *et al*, 2005).

## 5.12 USE OF POOLED SAMPLES

Often, to reduce cost, or when individual results are not needed or individual samples are not available, specimens from a number of animals might be pooled and tested as one sample. Such an approach is most efficient when  $P$  is low. Some of the issues that may affect the  $Se$  and  $Sp$  of a pooled sample (designated  $PISe$  and  $PISp$  respectively) are: homogeneity of mixing (more likely to be a problem with fecal samples than serum samples), whether individual samples are pooled at the laboratory or in the field (*eg* multiple swabs in a single tube of transport medium), the effects of dilution of the substance being tested for (perhaps to below the laboratory  $Se$ ), the characteristics of the animals whose samples are going into the pool, and the increased possibility of having extraneous cross-reacting substances added to the pool because of the inclusion of material from more animals.

The dilution effect has been demonstrated in several studies. For example, when pooling of saline from ear notch samples, the  $PISe$  of an antigen capture ELISA for bovine viral diarrhoea virus was shown to fall from 99% to 72% as the pool size increased from 2 to 5 (Cleveland *et al*, 2006). A similar effect was demonstrated for pooled swab samples for *Salmonella* from swine carcasses (Sørensen *et al*, 2007). A large effect of the characteristics (level of shedding) on  $PLSe$  was observed in a study into the use of pooled samples for Map (van Schaik *et al*, 2003). Two simulation studies evaluating the effects of various pool and test characteristics on  $PISe$  and  $PISp$ , along with the economic consequences of using pooled samples, have recently been published (Jordan, 2005; Muñoz-Zanzi *et al*, 2006).

An Internet-based program for estimating disease prevalence from pooled samples under a variety of conditions (*eg* known vs unknown  $Se$  and  $Sp$  of test) is available (<http://www.ausvet.com.au/pprev/>). Both frequentist and Bayesian methods of estimating prevalence from pooled samples used in the program have been reviewed (Cowling *et al*, 1999). This program was recently used in a study of Map in sheep (Toribio & Sergeant, 2007).

### 5.12.1 Pooled testing and HSe

Christensen and Gardner (2000) showed that  $HSe$  based on  $r$  pooled samples, each containing material from  $m$  animals, and assuming homogeneous mixing and no dilution effect is:

$$HSe = 1 - [(1 - (1 - P)^m)(1 - PISe) + (1 - P)^m PISp]^r \quad \text{Eq 5.40}$$

**Example 5.15 Estimating  $HSe$  and  $HSp$  from pooled specimens**

We can suppose that we are going to test herds for Map using pooled fecal culture. Fecal culture has an estimated  $PISe$  of 0.647 and  $PISp$  of 0.981. Suppose we wish to pool fecal samples from 5 cows together and we will use 6 pooled samples per herd. Hence,  $m = 5$  and  $r = 6$ .

If the herd is  $D$ -, then the herd  $Sp$  based on the pooled sample (assuming homogenous mixing) is:

$$(HSp) = (PISp)^r = (Sp^m)^r = (0.981^5)^6 = 0.562$$

If the herd is infected with a true prevalence of 12%, and assuming no dilution effect, then  $HSe$  is:

$$\begin{aligned} HSe &= 1 - [(1 - (1-0.12)^5)(1-0.647) + (1-0.12)^5 * (0.981)^5]^6 \\ &= 1 - [(1 - 0.528)(0.353) + 0.528 * 0.909]^6 \\ &= 1 - [0.167 + 0.480]^6 \\ &= 1 - 0.073 = 0.927 \end{aligned}$$

As with individual testing, the  $Se$  at the herd level is increased by testing more animals through the use of pooled samples but the  $Sp$  at the herd level is decreased. One could compare the 2 approaches ignoring costs and then add the cost of information to the final decision-making process.

If the herd is  $D$ -, then the herd  $Sp$  based on the pooled sample ( $HSp$ ) is  $(PISp)^r$ , and if no clustering occurs within pools,  $PISp=Sp^m$ . Thus, if pooled testing is performed on a number of assumed  $D$ - herds, then  $HAP=1-HSp=1-(PISp)^r$  which allows one to solve for the unknown  $PISp$ . Similarly, because  $Sp=PISp^{1/m}$ , increasing  $r$  or  $m$  increases the  $HSe$  and decreases  $HSp$  in the same manner as increasing  $n$  when testing individuals within a herd. The optimal choice of  $r$  and  $m$  should be investigated on a case-by-case basis. An example of estimating  $HSe$  and  $HSp$  based on pooled specimens is shown in Example 5.15.

**REFERENCES**

- Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 2004; 60: 427-35.
- Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med* 1999; 18: 2987-3003.
- Arunvipas P, VanLeeuwen J, Dohoo I, Keefe G. Evaluation of the reliability and repeatability of automated milk urea nitrogen testing. *Can J Vet Res* 2003; 67: 60-3.
- Barnhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 2002; 58: 1020-7.
- Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. *J Biopharm Stat* 2007; 17: 529-69.
- Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307-10.
- Branscum AJ, Gardner IA, Johnson WO. Estimation of diagnostic-test sensitivity and

- specificity through Bayesian modeling. *Prev Vet Med* 2005; 68: 145-63.
- Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology* 1996; 7: 199-202.
- Brusco MJ, Stahl S, Steinley D. An implicit enumeration method for an exact test of weighted kappa. *Br J Math Stat Psychol* 2007; 60: 377-93.
- Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol* 1993; 46: 423-9.
- Choi BC. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *Am J Epidemiol* 1998; 148: 1127-32.
- Choi Y, Johnson WO, Thurmond MC. Diagnosis using predictive probabilities without cut-offs. *Stat Med* 2006; 25: 699-717.
- Christensen J, Gardner I. Herd-level interpretation of test results for epidemiologic studies of animal diseases *Prev Vet Med* 2000; 45: 83-106.
- Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990; 43: 551-8.
- Cleveland SM, Salman MD, Van Campen H. Assessment of a bovine viral diarrhea virus antigen capture ELISA and a microtiter virus isolation ELISA using pooled ear notch and serum samples. *J Vet Diagn Invest* 2006; 18: 395-8.
- Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; 20: 37-46.
- Connell F, Koepsell T. Measures of gain in certainty from a diagnostic test *Am J Epidemiol*. 1985; 121: 744-53.
- Cook RJ. Kappa and its dependence on marginal rates. In: *Encyclopedia of Biostatistics*, 2nd Ed. J Wiley & Sons: New York; 2007.
- Coughlin S, Trock B, Criqui M, Pickle L, Browner D, Tefft M. The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test *J Clin Epidemiol* 1992; 45: 1-7.
- Cowling DW, Gardner IA, Johnson WO. Comparison of methods for estimation of individual-level prevalence based on pooled samples. *Prev Vet Med* 1999; 39: 211-25.
- de Vet H. Observer reliability and agreement. In: *Encyclopedia of Biostatistics*, 2nd Ed. J Wiley & Sons: New York; 2007.
- de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol* 2006; 59: 1033-9.
- Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001; 57: 158-67.
- Dohoo I, Wright P, Ruckerbauer G, Samagh B, Robertson F, Forbes L. A comparison of five serological tests for bovine brucellosis. *Can J Vet Res* 1986; 50: 485-93.
- Engel B, Buist W, Orsel K, Dekker A, de Clercq K, Grazioli S, van Roermund H. A Bayesian

- evaluation of six diagnostic tests for foot-and-mouth disease for vaccinated and non-vaccinated cattle *Prev Vet Med* 2008; 86: 124-38.
- Enøe C, Georgiadis MP, Johnson WO. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev Vet Med* 2000; 45: 61-81.
- Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Stat Med* 2002; 21: 3093-106.
- Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990; 43: 543-9.
- Fleiss JL, Levin B, Paik MC. *Statistical methods for rates and proportions*. 3rd Ed. J Wiley & Sons: New York; 2003.
- Gardner IA, Greiner M. Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. *Vet Clin Path.* 2006; 35: 8-17.
- Gardner IA, Stryhn H, Lind P, Collins MT. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Prev Vet Med* 2000; 45: 107-22.
- Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. *Am J Epidemiol* 1966; 83: 593-602.
- Georgiadis MP, Johnson WO, Gardner IA. Sample size determination for estimation of the accuracy of two conditionally independent tests in the absence of a gold standard. *Prev Vet Med* 2005; 71: 1-10.
- Georgiadis MP, Johnson WO, Gardner IA, Singh R. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Appl Stat* 2003; 52: 63-76.
- Giard RW, Hermans J. The diagnostic information of tests for the detection of cancer: the usefulness of the likelihood ratio concept *Eur J Cancer.* 1996; 32A: 2042-8.
- Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. *J Clin Epidemiol* 1993; 46: 1055-62.
- Greiner M. Analysis of diagnostic test evaluation data using survey statistics. In: *Proceedings of the 10th Symposium of the International Society for Veterinary Epidemiology and Economics*, Vina Del Mar, Chile. 2003. p. 243.
- Greiner M, Gardner I. Epidemiologic issues in the validation of veterinary diagnostic tests. *Prev Vet Med* 2000a; 45: 3-22.
- Greiner M, Gardner I. Application of diagnostic tests in veterinary epidemiologic studies. *Prev Vet Med* 2000b; 45: 43-59.
- Greiner M, Pfeiffer D, Smith R. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev.Vet Med.* 2000; 45: 23-41.
- Haber MJ, Barnhart HX. A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements. *Stat*



Methods Med Res. 2008; 17: 151-65.

Hajian-Tilaki KO, Hanley JA. Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data. *Acad Radiol.* 2002; 9: 1278-85.

Hanson TE, Johnson WO, Gardner IA. Log-linear and logistic modeling of dependence among diagnostic tests. *Prev Vet Med* 2000; 45: 123-37.

Hui SL, Walter SD. Estimating the error rates of diagnostic tests *Biometrics* 1980; 36: 167-71.

Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards *Stat Methods Med Res* 1998; 7: 354-70.

Jacobson RH. Validation of serological assays for diagnosis of infectious diseases *Rev Sci Tech* 1998; 17: 469-526.

Jordan D. Simulating the sensitivity of pooled-sample herd tests for fecal *Salmonella* in cattle. *Prev Vet Med* 2005; 70: 59-73.

Jordan D, McEwen SA. Herd-level test performance based on uncertain estimates of individual test performance, individual true prevalence and herd true prevalence. *Prev Vet Med* 1998; 36: 187-209.

Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol* 1995; 141: 263-72.

King TS, Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. *Stat Med* 2001; 20: 2131-47.

King TS, Chinchilli VM, Carrasco JL. A repeated measures concordance correlation coefficient. *Stat Med* 2007; 26: 3095-113.

Kostoulas P, Leontides L, Enøe C, Billinis C, Florou M, Sofia M. Bayesian estimation of sensitivity and specificity of serum ELISA and faecal culture for diagnosis of paratuberculosis in Greek dairy sheep and goats. *Prev Vet Med* 2006; 76: 56-73.

Lachenbruch PA. McNemar Test. In: *Encyclopedia of Biostatistics*, 2nd Ed. J Wiley & Sons: New York; 2007.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-74.

Lin LI. A concordance correlation coefficient to evaluate reproducibility *Biometrics* 1989; 45: 255-68.

Lin LI. A note on the concordance correlation coefficient *Biometrics* 2000; 56: 324-5.

Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987; 126: 161-9.

Martin S, Shoukri M, Thoburn M. Evaluating the health status of herds based on tests applied to individuals. *Prev Vet Med* 1992; 14: 33-44.

- McDowell I, Newell C. *Measuring Health. A guide to Rating Scales and Questionnaires.* Oxford Univ. Press: Oxford; 1996.
- McNamee R. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. *Stat Med* 2002; 21: 3609-25.
- McNemar Q. Note on the sampling error of the idfference between correlated proportions *Psychometrika* 1947; 12: 153-7.
- Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case-control studies. *Stat Med* 2007; 26: 2170-83.
- Miller W. Bias in discrepant analysis: when two wrongs don't make a right. *J Clin Epidemiol* 1998; 51: 219-31.
- Muñoz-Zanzi C, Thurmond M, Hietala S, Johnson W. Factors affecting sensitivity and specificity of pooled-sample testing for diagnosis of low prevalence infections. *Prev Vet Med* 2006; 74: 309-22.
- Nam J. Comparison of validity of assessment methods using indices of adjusted agreement. *Stat Med* 2007; 26: 620-32.
- Nerette P, Dohoo I, Hammell L. Estimation of specificity and sensitivity of three diagnostic tests for infectious salmon anaemia virus in the absence of a gold standard. *J Fish Dis* 2005a; 28: 89-99.
- Nerette P, Dohoo I, Hammell L, Gagne N, Barbash P, Maclean S, Yason C. Estimation of the repeatability and reproducibility of three diagnostic tests for infectious salmon anaemia virus. *J Fish Dis* 2005b; 28: 101-10.
- Nerette P, Stryhn H, Dohoo I, Hammell K. Using pseudogold standareds and latent class analysis in combination to evaluate the accuracy of three diagnostic tests. *Prev Vet Med* 2008; 85: 207-25.
- Nielsen SS, Grønbaek C, Agger JF, Houe H. Maximum-likelihood estimation of sensitivity and specificity of ELISAs and faecal culture for diagnosis of paratuberculosis *Prev Vet Med* 2002; 53: 191-204.
- Norby B, Bartlett PC, Grooms DL, Kaneene JB, Bruning-Fann CS. Use of simulation modeling to estimate herd-level sensitivity, specificity, and predictive values of diagnostic tests for detection of tuberculosis in cattle. *Am J Vet Res* 2005; 66: 1285-91.
- Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press: Oxford; 2003.
- Pouillot R, Gerbier G, Gardner I. "TAGS", a program for the evaluation of test accuracy in the absence of a gold standard. *Prev Vet Med* 2002; 53: 67-81.
- Reichenheim ME. Two-graph receiver operating characteristic *Stata J* 2002; 2: 351-7.
- Reichenheim ME. Confidence intervals for the kappa statistic *Stata J* 2004; 4: 241-8.
- Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol* 1978; 107: 71-6.

- Saah A, Hoover D. "Sensitivity" and "specificity" reconsidered: the meaning of these terms in analytical and diagnostic settings *Ann Intern Med*. 1997; 126: 91-4.
- Sargeant JM, Martin SW. The dependence of kappa on attribute prevalence when assessing the repeatability of questionnaire data. *Prev Vet Med* 1998; 34: 115-23.
- Sørensen LL, Wachmann H, Alban L. Estimation of Salmonella prevalence on individual-level based upon pooled swab samples from swine carcasses. *Vet Microbiol* 2007; 119: 213-20.
- Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. *Arch Gen Psychiatry*. 1985; 42: 725-8.
- Streiner DL, Norman GR. "Precision" and "accuracy": two terms that are neither. *J Clin Epidemiol* 2006; 59: 327-30.
- Suzuki S. Conditional relative odds ratio and comparison of accuracy of diagnostic tests based on 2 x 2 tables. *J Epidemiol* 2006; 16: 145-53.
- Thomsen PT, Baadsgaard NP. Intra- and inter-observer agreement of a protocol for clinical examination of dairy cows. *Prev Vet Med* 2006; 75: 133-9.
- Toft N, Jørgensen E, Højsgaard S. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. *Prev Vet Med* 2005; 68: 19-33.
- Toribio JLML, Sergeant ESG. A comparison of methods to estimate the prevalence of ovine Johne's infection from pooled faecal samples. *Aust Vet J* 2007; 85: 317-24.
- Vach W. The dependence of Cohen's kappa on the prevalence does not matter. *J Clin Epidemiol* 2005; 58: 655-61.
- van Schaik G, Stehman SM, Schukken YH, Rossiter CR, Shin SJ. Pooled fecal culture sampling for *Mycobacterium avium* subsp. *paratuberculosis* at different herd sizes and prevalence. *J Vet Diagn Invest* 2003; 15: 233-41.
- Walter SD. The problem of imperfect reference standards. *J Clin Epidemiol* 2005a; 58: 649-50.
- Walter SD. The partial area under the summary ROC curve. *Stat Med* 2005b; 24: 2025-40.
- Wan S, Zhang B. Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. *Stat Med* 2007; 26: 2565-86.
- Woodard DB, Gelfand AE, Barlow WE, Elmore JG. Performance assessment for radiologists interpreting screening mammography. *Stat Med* 2007; 26: 1532-51.
- World Organisation for Animal Health (OIE). *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals*, 5th Ed. Paris: OIE; 2004.
- Wruck LM, Yiannoutsos CT, Hughes MD. A sequential design to estimate sensitivity and specificity of a diagnostic or screening test. *Stat Med* 2006; 25: 3458-73.
- Zou G. From diagnostic accuracy to accurate diagnosis: interpreting a test result with confidence *Med Decis Making* 2004; 24: 313-8.

