Principles of diagnostic testing of animal populations

Peter Davies BVSc, PhD
University of Minnesota, St. Paul, Minnesota

Along with almost everything else in the swine industry, the way veterinarians apply and interpret diagnostic tests has evolved as production systems have changed. Valid interpretation of diagnostic results requires integration of clinical, epidemiological and laboratory observations, and consideration of potential sources of error. Beyond their traditional application in outbreak investigation and disease diagnosis, diagnostic tests are increasingly used to quantify disease in defined populations in order to support decisions regarding population health management. The optimal strategy for population testing, including the tests selected and the numbers of animals sampled, is a function of the purpose of testing (Table 1).1 Fundamental objectives are to establish the disease status (positive or negative) of populations (groups, sites, studs, pods, companies, regions, countries), or to quantify disease trends in time or space. Demographic changes (particularly increased herd sizes) and persistent problems with PRRS have altered the needs for, and feasibility of, more intensive use of herd level diagnostics.2 Veterinarians must have a sound theoretical understanding of the relationships between diagnostic test performance, sampling theory, and disease prevalence that underpin population testing. This paper reviews some core principles that are essential for veterinarians to be able to meet these challenges. More detailed discussions of diagnostic testing of swine populations, and approaches to disease surveillance in swine are available elsewhere.1,3

Table 1: Purposes of diagnostic testing in veterinary medicine (adapted from Gardner and Blanchett, 20051)

Detect agents responsible for disease outbreaks or suboptimal production
Evaluate the infection/exposure status of individual pigs
Determine whether a herd is infected with a pathogen
Determine age or production groups affected
Estimate the percentage of herds or pigs with antibodies to an infectious agent
Monitor a herd’s serologic response to vaccination
Monitor the progress and success of disease control and eradication programs
Fundamentals of population sampling for disease

When contemplating testing a population for disease, it is important to clarify the underlying purpose. Usually this is either detection of disease, or estimation of prevalence or incidence. It is important to define which of these is your primary objective, because

a) the underlying calculations of appropriate sample size are different, and
b) unbiased selection of animals is required for prevalence estimation, whereas deliberate bias of sampling towards high risk animals can be desirable if the aim is to detect disease

The selection of the number of animals to sample is driven by both requirements (accuracy and level of confidence; cost constraints) and assumptions (expected prevalence). For detection of disease, the factors that determine required sample size include the population size, the minimum expected prevalence of disease if it is present, and the level of confidence for detecting disease at this prevalence. User friendly software to perform these calculations is freely available on the internet (e.g. WinEpi at http://www.winepi.net/uk/index.htm). Although this program does assume perfect test performance, it is useful for quickly looking at relationships between sample size and probability of detection, and also for evaluating results.

Many veterinarians have been raised on the familiar sample size of 30 (adequate to detect at least one positive with 95% confidence if prevalence is 10% or more). However, it is important to remember that samples of 30 will be inadequate to reliably detect disease when prevalence is lower (Figure 1).

Figure 1: Expected numbers of infected animals in samples of 30 animals selected from populations with prevalence of 5% or 10%.

Note that when prevalence is 5%, more than 20% of the time samples of 30 from a large population will not include a single diseased animal. Another way to look at negative test results is the likely maximum prevalence. With 30 negative samples from a population of 1000,
assuming perfect test sensitivity you would be 95% confident that the prevalence in the population was below 9.5%. Alternatively, if you have a sample size of 10 animals, you would only be 95% that the prevalence in the herd was below 25.8%. Such a small sample size is therefore only informative when the expected prevalence is high (25% or greater).

Sampling to estimate disease prevalence is more frequently a research rather than a clinical objective. Similar to the case of detecting disease, key determinants of required sample size are population size, required confidence level, and expected prevalence. A fourth factor is the required precision, expressed as the allowable error in the estimate (e.g. within 5% or 10%). Because the variance of a proportion is greatest at 50%, relatively small samples are required to estimate prevalence when it is very high (>90%) or very low (Figure 2).

Figure 2: Effect of expected disease prevalence on the sample size required to estimate disease prevalence within 10% with 90% or 95% confidence (perfect test assumed).

In the absence of data on expected prevalence, a conservative approach is to use the sample size indicated for 50% prevalence. Again, these guidelines assume perfect test performance and may need to be modified in accordance with test performance parameters.

Diagnostic test performance at the individual level

Diagnosis is an imperfect process. An essential foundation of effective population testing is comprehension of the principles of diagnostic testing at the individual level. Minimally this means understanding the quantitative relationships between test performance parameters (sensitivity and specificity) and disease prevalence (or the ‘pre-test probability of disease’) as determinants of the predictive value of test results (i.e. the probability that the disease status of the animal or group corresponds with the test result). These concepts should be thoroughly
familiar to swine veterinarians, and thus are reviewed only briefly as background for discussion of testing at the aggregate (herd or population) level. Although test data may be dichotomous (e.g. culture positive or negative), ordinal (e.g. serological titers), or continuous (e.g. OD or S/P ratio), most test results are ultimately interpreted as positive (indicating disease or exposure) or negative (indicating no disease or exposure). The sensitivity (Se) and specificity (Sp) of tests are defined as the probability of the test correctly classifying diseased and healthy individuals respectively (Figure 3). That is, given the status of the animal (diseased or not), what is the probability that the test result is correct. While these are important parameters for evaluating test performance, typically the question the veterinarian wants to answer is the predictive value of test results – i.e. given a test result (positive or negative), what is the probability that the disease is present (or absent). Thus positive predictive value (PPV) is the probability that a test positive animal is diseased, and negative predictive value (NPV) is the probability that a test negative animal is not diseased.

Figure 3: Basic measures of individual diagnostic test performance

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Disease Status</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present</td>
<td>True positive</td>
<td>False Positive</td>
<td>False negative</td>
<td>True negative</td>
</tr>
<tr>
<td></td>
<td>Absent</td>
<td></td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

Sensitivity = probability of diseased animal testing positive = \( \frac{A}{A+C} \)
Specificity = probability of non-diseased animal testing negative = \( \frac{D}{B+D} \)

False negative fraction = probability of diseased animal testing negative = \( \frac{C}{A+C} \)
False positive fraction = probability of diseased animal testing positive = \( \frac{B}{B+D} \)

Positive predictive value = probability a test positive animal is diseased = \( \frac{A}{A+B} \)
Negative predictive value = probability a test negative animal is not diseased = \( \frac{C}{C+D} \)

While sensitivity, specificity and other measures derived from them (e.g., ROC curves and likelihood ratios) are essentially independent of disease prevalence, test predictive values are strongly influenced by prevalence (Figure 4). More generally, it is the ‘pretest probability of disease’ (in this case prevalence) that influences the interpretation of a given test result. When the pre-test probability of disease is low (i.e. a rare or unlikely disease), the predictive value of a positive test will be low unless the test specificity is extremely high.
Although these relationships are straightforward in theory, their application in practice is not. Almost invariably, there is considerable uncertainty (well recognized by veterinarians) about the pre-test probability of disease. However, estimates of test Se and Sp are also uncertain. We tend to view published (or stated) values of Se and SP as both fixed and precise values. However, reliable estimates of test Se and SP are often not available for the target populations in which tests are applied. Estimates obtained from experimental disease models will typically overestimate test performance, and available data may not be representative of the target population (i.e. the population in which you are using the test). Available estimates rarely include confidence intervals to indicate the associated uncertainty. Furthermore, it is usual that test performance will vary with the stage of a disease process, yet this variability is rarely expressed in test performance measures. Ideally, veterinarians need precise and unbiased estimates of diagnostic test performance in the population that is under investigation and for the purpose the test is being employed. Most often, available data fall short of this standard, and we need to acknowledge and manage uncertainty about test performance in all diagnostic endeavors. When presented with figures on test performance parameters, we should look into information about how these estimates were derived.

Determinants of aggregate (herd) test performance

The concepts of Se, Sp and predictive values translate directly from the individual to the aggregate (population) scale. The sensitivity of a herd diagnostic test is the probability that the testing protocol will yield a positive result when an infected population is tested; herd test specificity is the probability that the testing protocol will yield a negative result when a disease-free population is tested. Herd PPV is the probability that disease is present in a population that tests positive; and herd NPV is the probability that disease is absent from a group that tests negative. Individual test performance measures (Se, Sp) are obviously key determinants of the performance of herd level tests. However, it is important to consider other determinants of herd test performance and their interrelationships. In the same way that selection of a cut-off (e.g. S/P
ratio) for an individual animal test involves a compromise between Se and Sp, design of population testing protocols shares the same dilemma. Table 2 shows the factors that combine mathematically to determine the performance of aggregate testing protocols.

Table 2: Determinants of Herd Sensitivity (Herd Se) and Herd Specificity (Herd Sp) in population testing.

<table>
<thead>
<tr>
<th>Within Herd Prevalence (p)</th>
<th>n tested</th>
<th>Individual test Se</th>
<th>Individual test Sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herd Se</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Herd Sp</td>
<td>- (=0)</td>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

Determinants of herd test specificity (HSp)

By definition, herd specificity applies only to disease free populations, and is the probability that the testing protocol will correctly classify the herd as negative. The complement of this (i.e., 1-HSp), is the probability of a false positive test. As all animals in a disease free population are negative, test sensitivity is irrelevant, and HSp is determined by only two factors: the number tested (n) and individual test specificity (Sp). The probability that all samples will be negative is simply (Sp)^n. Note that HSp declines exponentially as n increases (Figure 5).

Figure 5: Relationship between sample size (n) and herd test specificity for various values of individual test specificity (0.95, 0.98, 0.99, 0.999).
For this reason, unless Sp is extremely high (>0.99) testing large numbers of animals will usually generate some false positive results, which have to be anticipated and managed. This is most problematic when veterinarians are testing for diseases that are expected to occur at low prevalence within a population. In these situations, direct detection of the agent (e.g. by isolation or PCR), is likely to be more effective than testing for antibody.

**Determinants of herd test sensitivity (HSe)**

In the simplest scenario, a herd test will be deemed positive if any animal tested returns a positive test. However, if a diseased herd is tested, the observed proportion of positive tests (‘apparent prevalence’) may include both true positive and false positive animals. Apparent (not true) prevalence is the probability that an individual animal sampled from a population will be test positive. Mathematically, the relationship between true prevalence (TP) and apparent prevalence (AP) is:

\[
AP = TP*Se + (1-TP)(1-Sp)
\]

Thus the true prevalence of disease as well as individual Se and Sp will influence the probability of a positive herd test. To begin, let’s look at a simplified scenario (e.g. bacterial culture) in which test Sp is considered perfect (i.e. no false positive individual results). In this scenario, the second term in the above equation equals zero (as 1-Sp = 0), therefore:

\[
AP = TP * Se, \text{ and the probability of an animal testing negative is } 1 – (TP*Se)
\]

To obtain a positive herd test, we need to detect only one infected individual. If we test a sample of \(n\) animals, the probability that they will all test negative is \((1-AP)^n\). Consequently, the herd test sensitivity (probability that one or more will test positive) is \((1-(1-AP)^n)\). Figure 4 illustrates the impact of sample size on herd Se and herd Sp for a given test at a range of true prevalences.
Figure 6: Relationship between sample size (n) and herd test sensitivity for various values of true prevalence (2%; 5%; 10%; 25%). Individual test sensitivity is 0.95 and specificity is 0.98, and the change of herd test specificity with sample size is also shown.

Another useful concept in evaluating test results is likelihood ratios. A likelihood ratio is a ratio of the probability that an observed test result is found in a diseased animal (or herd) to the probability of that result in a disease negative animal (herd).

Likelihood ratio = \frac{\text{Probability of test result in a diseased individual}}{\text{Probability of test result in a non-disease individual}}

A ratio of 1 would indicate that a positive test was as equally likely to result from diseased (true positive) and non-diseased (false animal), and that the test would therefore provide no useful information. Likelihood ratios are usually evaluated in relation to individual tests, and are the ratio of test sensitivity to the false positive rate. In the context of a herd test, high likelihood ratios will occur when prevalence and Se are high, and low ratios will occur when prevalence and test specificity are low. At a true prevalence of disease were 2% with a test Se of 0.95 and Sp of 0.98, a false positive test is more likely to originate from a non-diseased than a diseased animal, but this ratio increases substantially as prevalence increases (Table 3).

Table 3: Likelihood ratios for a positive test in herd testing scenarios of 2, 5, 10, and 25% true prevalence with individual test sensitivity of 95% and specificity of 98%

<table>
<thead>
<tr>
<th>True Prevalence</th>
<th>P true positive</th>
<th>P false positive</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2%</td>
<td>0.019</td>
<td>0.0196</td>
<td>0.97</td>
</tr>
<tr>
<td>5%</td>
<td>0.0475</td>
<td>0.019</td>
<td>2.5</td>
</tr>
<tr>
<td>10%</td>
<td>0.095</td>
<td>0.018</td>
<td>5.3</td>
</tr>
<tr>
<td>25%</td>
<td>0.2375</td>
<td>0.015</td>
<td>15.8</td>
</tr>
</tbody>
</table>
In recent years, there have been considerable theoretical advances in diagnostic test interpretation in veterinary medicine, particularly at the aggregate. These have led to the development of some practical tools to facilitate quantitative approaches to the design and interpretation of population testing protocols. For example, Survey Toolbox (http://epitools.ausvet.com.au/content.php?page=SurveyToolbox) is a suite of software programs designed to support planning and analysis of animal health surveys. One component of this suite is an epidemiological probability calculator (FreeCalc) to support testing to evaluate freedom from disease. FreeCalc has two modules that enable calculation of sample sizes and analysis of test results for freedom from disease. It accounts for imperfections in diagnostic tests and population size, and requires inputs for test Se and specificity, population size, minimum expected prevalence of disease (if present), and allows the user to define acceptable error levels. More sophisticated tools are becoming available that extend this approach to incorporate uncertainty in test performance characteristics and expected prevalence. The latter publication explains freely available software (BayesFreeCalc) for evaluating the probability of freedom from disease allowing for uncertainty in input parameters.

Adjusting the cut-off (k) number of positives

In the same way that a cut-off for an ELISA test involves inherent compromise between individual Se and Sp, in herd testing one can arbitrarily alter the cut-off number of animals (i.e. number of positives required for a herd test to be deemed positive). Clearly, maximal sensitivity will be obtained if k = 1 (i.e. herd deemed positive if one animal tests positive), but as shown above specificity can be a problem. For cost reasons, there is an argument for limiting sample size and using k = 1 as a cut-off, however the limitations of this approach for low prevalence diseases should be clear. As with individual tests, the desire for high specificity vs. high sensitivity is a function of both the pre-test probability of disease and of the costs of diagnostic errors (termed ‘misclassification costs’). Where the costs of a false negative result are much greater than the costs of a false positive result, high sensitivity (i.e. a low cut off) is required. Conversely, if the cost of a false positive test is relatively high, higher specificity is desirable. These theoretical relationships can be expressed in the misclassification cost term (MCT):

$$MCT = (1-P)(1-Sp) + rP(1-Se)$$

Where r is the ratio of the cost of a false positive test to that of a false negative test, and P is the pretest probability of disease. The optimal cutoff is defined as the values of Se and Sp that minimize the MCT. Again, at a practical level, uncertainties regarding the pre-test probability of disease, HSe and HSp constrain the application of such quantitative approaches. However, as we increasingly employ protocols for routine population testing in large production systems, more accurate estimation of misclassification costs will be an important of optimizing the diagnostic process.

Interpretation of herd testing results
In summary, proper interpretation of herd tests relies on sound understanding of the determinants of herd test performance. Interpretation of herd tests is more complicated than for individual tests. Disease prevalence within herds is analogous to the ‘stage of disease’ in an individual test, and has major impact on herd testing performance. The imperfect specificity of an individual test is magnified when large numbers of individuals are tested. Veterinarians need to quantify the probability of false positive results in herd testing programs, and to establish protocols for managing them (e.g. retesting procedures). Imperfect sensitivity is to some extent less problematic (unless complete assurance of disease freedom is the goal) and can be compensated for by increasing sample size. The two decision thresholds for herd tests (the cut-off for the individual test; the cut-off of k positive animals) introduce some flexibility in selecting HSe and HSp by varying n and k (assuming estimates of individual Se and Sp are available). However, it may be impossible (with adequate confidence) to reliably discriminate between infected and non-infected herds when prevalence is low.

References
