CHAPTER 2.2.5.

STATISTICAL APPROACHES TO VALIDATION

INTRODUCTION

The WOAH Validation Recommendations in Section 2.2 Validation of diagnostic tests *of this* Terrestrial Manual *provide detailed information and examples in support of the WOAH Validation Standard that is published as Chapter 1.1.6* Validation of diagnostic assays for infectious diseases of terrestrial animals. *The Term "WOAH Validation Standard" in this chapter should be taken as referring to that chapter.*

The choice of statistical methods for analysis of test validation data from laboratory experiments and from evaluation of field-based samples depends on considerations such as the experimental design and sample selection (source, number of samples, number of replicates of tests, etc.). Specific guidance about the "best approach" should be made in consultation with a statistician and should be done during the design phase before validation studies commence.

Definitions of scales of measurement:

Binary (dichotomous): Either positive or negative because that is how the test result is presented, or positive/negative at a selected threshold (cut-off) value when results are measured on an ordinal or continuous scale.

Ordinal: Measured on a scale with discrete values where higher values typically indicate more analyte, e.g. serum virus neutralisation titres.

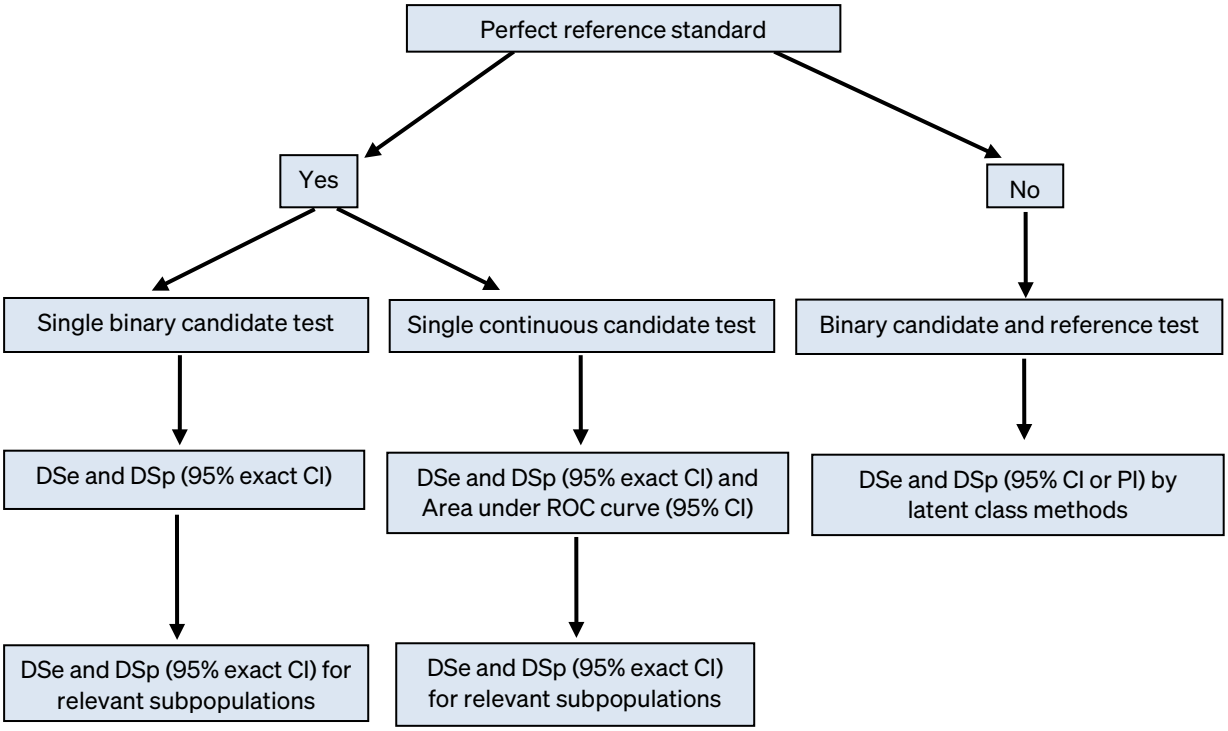
Continuous: An infinite number of measured values are theoretically possible, depending on the measurement system, e.g. optical density or per cent positive values in an enzyme-linked immunosorbent assay, and cycle threshold values from real-time polymerase chain reaction assays that are lower than the maximum number of cycles that are run in the assay.

For brevity, this annex considers commonly-used approaches for validation of a candidate test and hence, does not consider all statistical methods that might be used in practice. Methods are described to estimate the precision of an assay when repeated multiple times (repeatability and reproducibility), analytical characteristics (analytical sensitivity and specificity) and diagnostic characteristics (e.g. diagnostic sensitivity [DSe] and specificity [DSp], and area under the receiver-operating characteristic curve of an assay) used to detect an analyte in individual animals. Similar principles apply when tests are used to detect the same analyte in naturally or artificially-created sample pools from animals in aggregates (e.g. herds or flocks). In this case, the epidemiological unit is the aggregate rather than individual animals.

Statistical methods differ depending on whether a single or multiple tests are evaluated, their scales of measurement (binary, ordinal, or continuous), whether independent or dependent (paired) samples are used, and whether there is a perfectly accurate reference standard (often termed a gold standard) for comparison (Wilks, 2001). Flow charts to guide selection of statistical methods for evaluation of diagnostic accuracy measures such as sensitivity and specificity are in Figures 1 and 2.

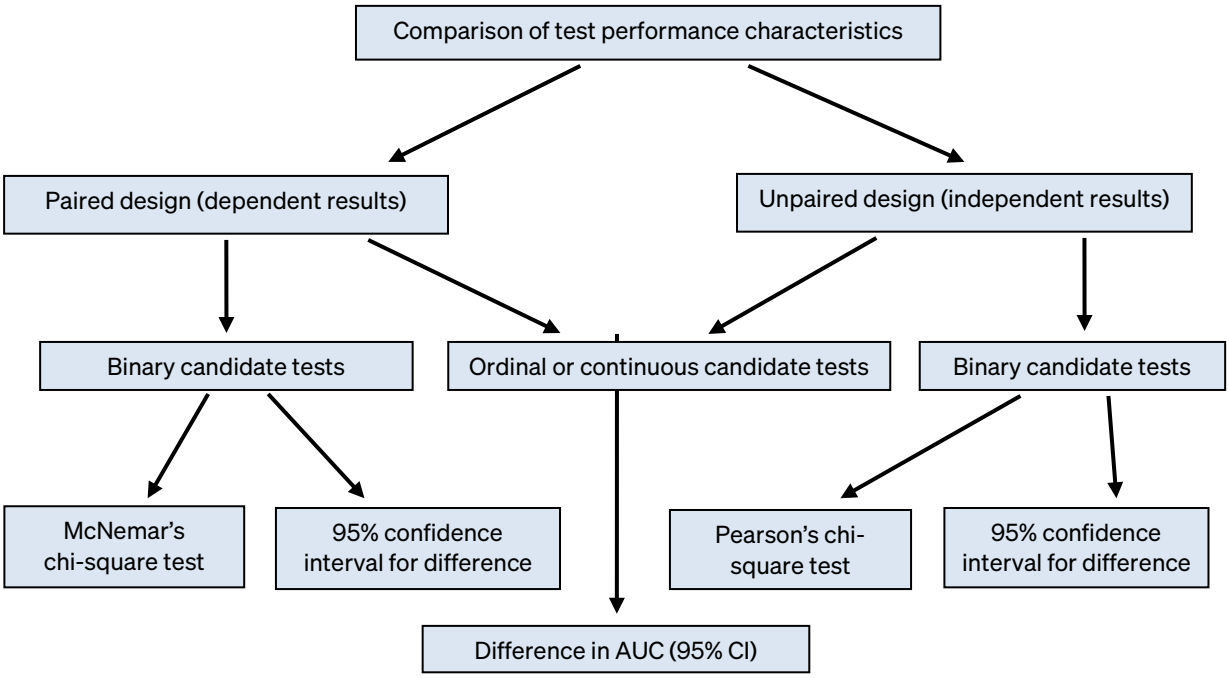
*The adequacy of the design of the study and statistical analysis may not always be reflected in the quality of reporting in scientific publications and hence, test developers and evaluators are encouraged to follow the STARD (**Sta**ndards for **R**eporting of **D**iagnostic Accuracy) checklist (Bossuyt* et al., *2003) to ensure complete reporting of all relevant information in validation studies of infectious diseases in animals.*

For guidance on analysis of measurement uncertainty data and for data for methods comparison studies refer to Chapters 2.2.4 and 2.2.8, respectively.



Abbreviations: DSe = diagnostic sensitivity; DSp = diagnostic specificity; ROC = receiver operating characteristic; CI = confidence interval; PI = probability interval.

Fig. 1. Flow chart for suggested methods of statistical analysis when a single candidate test is evaluated with and without a perfect reference standard.



Abbreviation: AUC = area under the receiver operating characteristic curve

Fig. 2. Flow chart for suggested methods of statistical analysis when sensitivities (DSe) and specificities (DSp), and the AUC of multiple tests, are evaluated with a perfect reference standard. Ordinal and continuous data should be analysed in their original form and as binary results at the recommended thresholds. Analyses should be done for both DSe and DSp where these data are available.

A. ASSAY REPEATABILITY WITHIN A SINGLE LABORATORY

An assessment of within-laboratory repeatability of an assay (often termed *precision* when the measurement is on a continuous scale) requires that a minimum of three samples having analyte concentrations within the operating range of the assay are tested in replicate by a single operator using a single test-kit lot or batch. Typically these runs will be on the same day but separate days are also possible. Use of three or four replicates of a sample rather than two is encouraged because it better captures the inherent variability in within-run assay results. Because of cost considerations, use of more than two replicates may not be feasible for all assay types (e.g. nucleic-acid detection). As described in the WOAH Validation Standard, between-run variation can be evaluated in multiple runs involving two or more operators on multiple days. The following two sections describe approaches to analysis of continuous and binary data for assay repeatability.

1. Continuous outcomes

For continuous outcomes, the simplest approach is to estimate the standard deviation (SD) of replicates of a set of samples representing the operating range of the assay. These results initially should be evaluated in a scatter plot or diagram of the mean of the replicates plotted against the SD. For assays, where the SD is proportional to the mean, the within-sample coefficient of variation (CV) is often calculated. CV is often used even when proportionality does not exist. In this case, the CV should be reported for levels of the target analyte (e.g. low, moderate and high). This is necessary because it is a common finding that CV is often larger when the concentration of target analyte is low. In general, an estimate of uncertainty in the CV values (e.g. 95% confidence interval [CI]) should also be calculated. Where the CV values are fairly constant over the range of test values, this can be done using the results of all samples. Where CV differs according to analyte concentration, separate 95% CI should be calculated for each analyte category based on the number of samples tested at each level. Methods for CI calculation for CV and the difference in two CV for normal data are described in Donner & Zou (2012).

$$CV = \frac{\text{SD of replicates}}{\text{Mean of replicates}}$$

where: CV = Coefficient of variation
SD = Standard deviation

If the experimental design includes evaluation of multiple factors such as different operators and run days, other approaches such as variance component models (mixed models) may be needed should the goal be to decompose the variation into the sum of several components that can be readily interpreted. Variance components models can also be used for reproducibility data (see Section B).

2. Binary outcomes

In general, quantitative results should be used for evaluation of assay precision when data are available in that form, even though results might be dichtomotised for reporting purposes. For inherently binary tests which yield positive or negative results, the kappa statistic can be used to quantify the agreement of test results beyond chance. Kappa ranges from 0 (no agreement beyond chance) to 1 (perfect agreement beyond chance) but there is much conjecture about how kappa values should be interpreted (Fleiss *et al*., 2003; Landis & Koch 1977). Better agreement is typically expected when test results are well away from the cut-off points and hence, some samples with intermediate/suspicious values should be tested to avoid overly optimistic assessments of agreement. A weighted version of kappa for ordinal results (e.g. negative, suspicious, and positive) can be used to recognise that a large discrepancy (e.g. two category difference) is more serious than a smaller discrepancy (e.g. one category difference). Ninety-five per cent CI should be reported for unweighted or weighted estimates of kappa (Fleiss *et al.,* 2003).

Table 1. Examples of Kappa calculations for binary outcomes

Example 1: Kappa calculation based on repeated test results classified as positive or negative

Test result	***Positive***	***Negative***
Positive	90	5
Negative	10	95
	100	100

Kappa = 0.85 (95% CI = 0.78 to 0.92)

Example 2: Kappa calculation based on repeated test results classified into three categories (positive, suspicious, or negative)

Test result	*Positive*	*Suspicious*	*Negative*
Positive	80	10	10
Suspicious	15	75	10
Negative	5	15	80
	100	100	100

Kappa = 0.68 (95% CI = 0.61 to 0.75). Weighted kappa = 0.70. (95% CI = 0.61 to 0.79)

B. ASSAY REPRODUCIBILITY AMONG LABORATORIES

Assay precision will vary according to routine implementation, e.g. different operators, different test sites, using different kit lots, or on different days. Most commonly, the term *reproducibility* is applied to assessment of precision of the selected assay in multiple laboratories. Factors held constant should be described to allow interpretation of results in context to the actual testing situation. Reproducibility studies can be done independently of or in association with repeatability studies but should be done in a blinded fashion. As suggested in the WOAH Validation Standard, at least three laboratories should test a minimum of 20 samples with identical aliquots going to each laboratory.

Intraclass correlation coefficient represents the similarity or correlation of any two measurements made on the same sample. The ICC takes values between 0 and 1 with values closes to 1 indicating minimal measurement error. Conversely, values close to 0 indicate a large amount of measurement error.

Statistical methods for analysis of studies of assay reproducibility among laboratories are similar to those used for assessment of within-laboratory repeatability. However, as part of an among laboratory study, it might be considered important to assess and rank variability in test results from multiple sources (often termed a class). For example, if a study was designed to test an assay in three laboratories each using two highly-trained technicians and running the samples in duplicate on two kit lots, each test sample would be tested 24 times. The selected factors (laboratory, technician, kit lot, replicate result) can be considered to be fixed or random depending on how they are selected and whether they are representative of the target population. For this study design, variance components can be estimated for each class (example is Dargatz *et al.,* 2004) and the intraclass (intracluster) correlation coefficient (ICC) can be estimated as a measure of the similarity of sample results (Bartlett & Frost, 2008).

1. When a technical modification of the test method is made

After an assay has been validated for use in a controlled laboratory environment, it may be considered for use in a very different environment (such as a pen-side application). Because of the more extreme changes, for example severe temperature fluctuations which often occur at pen-side, it would be expected that the two tests might behave very differently in their different environments. In fact, rather than random measurement error which applies to the assessment of within or among laboratory measurement error, it is anticipated that the values in such a study likely would be interpreted as a systematic measurement error, which would be the case if the values provide an over- or under-estimate of the true value. For the example of a test run on split samples pen-side and in a laboratory, the mean of the differences between the pen-side value and the within laboratory value (true value) for the same sample should be reported with a 95% CI. If the 95% CI excludes zero, there is evidence of systematic deviation of test results when used pen-side compared within the laboratory. When such a systematic deviation in test results occurs, the pen-side test results are not comparable with those from the laboratory-based validated assay. To validate the pen-side assay, either it is subjected to a “technical modification” that is then evaluated in a methods comparison study (see Chapter 2.2.8) or a full re-validation of the pen-side application is required.

Similar approaches can be used to assess method changes within a laboratory to determine whether there is systematic or random variation in the results.

Example: The following unpublished data were obtained comparing two extraction methods (old and new on split samples) on cycle threshold (CT) values for a real-time polymerase chain reaction (PCR) for bluetongue. The data (n=10) represent means of sample duplicates.

Old method: 25.6, 24.5, 21.3, 26.8, 25.2, 30.2, 31.2, 32.8, 31.8, 34.9

New method: 23.1, 21.0, 18.2, 25.2, 24.7, 28.6, 30.4, 32.2, 31.3, 34.7

The mean difference between the two methods (old minus new) was –1.49 (95% CI = –2.33 to –0.64) with a two-tailed probability of p=0.003. Because the 95% CI excludes zero, this indicates a systematically lower CT value when the new extraction method is used. A Bland–Altman plot (Bland & Altman, 1999; Fig. 3) can be used to graphically depict how the difference changes as a function of the mean value of the old and new method. For these data, the difference appears to decrease for higher CT values but the sample size is small.

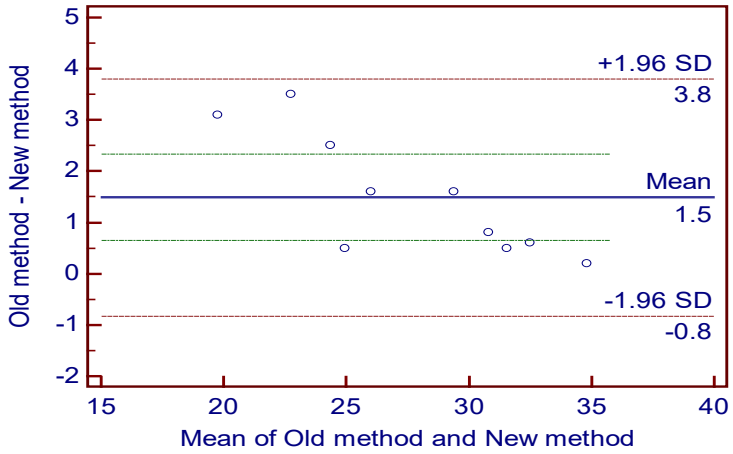


Fig. 3. Bland-Altman plot of mean difference (y axis) in CT values as a function of the mean value of the old method and new method (n=10).

C. ANALYTICAL SENSITIVITY (ASE, SYNONYM = LIMIT OF DETECTION: LOD)

Analytical sensitivity can be estimated using a dilution-to-extinction (DTE) experiment in which serial dilutions of a known quantified amount of target analyte are made into the appropriate sample matrix. This known quantified amount might be from an in-house or national/international reference standard or a field sample whose analyte concentration has been determined. Parallel runs of a comparison standard can be done but are not essential, unless the study is one in which a minor change in a validated assay is being compared with the original validated assay. The DTE approach can be used if the analyte is measured qualitatively or quantitatively. In the latter case, the test result is reclassified as positive or negative.

The approach to analysis of LOD data depends on the experimental design. For example, suppose that a study was done in which 10^8 colony-forming units (CFU) of a bacterium were spiked into 10 g faeces to achieve a concentration of 10^7 CFU/g. This sample was then diluted in tenfold serial dilutions to 10^1 CFU/g. The experiment was repeated three times. If all replicates at 10^3 CFU/g were detected but none at 10^2 CFU/g, the LOD could be conservatively estimated as 10^3 CFU/g. If a precise estimate of the LOD were needed, a second stage experiment could be designed to determine the LOD with a greater certainty using a series of finer dilutions, e.g. twofold, encompassing the interval between 100% detection and 0% detection identified in the first experiment. The LOD endpoint often is chosen to be 95%; in an experiment with 20 replicates, this corresponds to the dilution where 19 replicates for analyte were positive. The important point is that the chosen probability point for the LOD (whether 95%. 50% of another value) should be specified and used consistently if results of multiple tests are being compared. The LOD can be estimated using the Spearman-Kärber non-parametric approach, or by logistic regression or probit analysis. The greater the number of replicates for each dilution, the more precise the estimate of LOD.

Example: Guthrie *et al.* (2013) made a twofold dilution series of an AHSV-positive horse blood (10^{-3} dilution), which covered the non-linear range of the assay. The extraction was repeated 25 times and samples were tested by

AHSV real-time PCR. The real-time PCR results for the 15 dilution points were used in a probit analysis to calculate the 95% LOD (i.e. input concentration giving a positive real-time real-time PCR result in 95% of the replicates (Burns & Valdivia, 2008). The 95% LOD was estimated to be at a dilution of 3.02×10^{-6}, as shown in Figure 4, and corresponded to a quantification cycle of 35.71 in the real-time PCR. CI for the estimate were not reported.

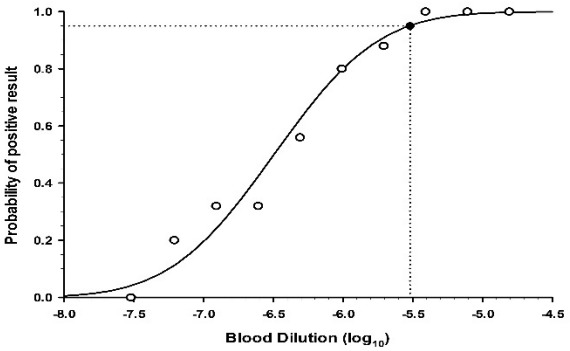


Fig. 4. Estimated 95% limit of detection of AHSV in horse blood (*log_{10}*), which is shown by the dashed line.

D. ANALYTICAL SPECIFICITY (ASP)

Analytical specificity can be described in at least three distinctive ways: selectivity, exclusivity (synonym: cross-reaction profile), and inclusivity (as described in the WOAH Validation Standard). The latter two measures should be reported on a per lineage, isolate, species or genus basis, as appropriate for the target analyte and the intended purpose of the test. For screening tests, a broader and more inclusive specificity is required than for a confirmatory test that may distinguish between isolates that, for instance, vary in pathogenicity. Because the choice of related organisms is subjective and often dependent on the types and numbers of samples, the exclusivity result should be reported qualitatively, e.g. the percentage of related agents that cross-reacted in the assay with a listing of potential cross-reacting agents that were evaluated. Similarly, inclusivity is reported as a percentage of the serovars, strains, genera and species detected by the assay, as appropriate for the target analyte.

E. DIAGNOSTIC PERFORMANCE OF THE ASSAY

Diagnostic performance of an assay is mostly commonly measured as sensitivity (DSe) or specificity (DSp) or a combined measure of DSe and DSp such as the likelihood ratio of a positive of negative result. Likelihood ratios for intervals of test results can also be calculated when it is important to retain information on the magnitude of the test result rather than use it in a dichotomised form. For more information on use and calculation of likelihood ratios, see Gardner & Greiner (2006) and Gardner *et al.* (2010). The latter paper includes an example for porcine toxoplasmosis with CI calculations by two methods.

Statistical uncertainty about diagnostic performance parameters, e.g. DSe and DSp, should be presented as confidence intervals (CI). Typically, a 95% CI is used and its width (precision of the estimated value) depends strongly on the sample size used for parameter estimation. Exact CI are preferred to normal approximations because they avoid upper limits that exceed 100%.

DSe and DSp can be estimated when the reference or comparison method is perfectly sensitive and specific or when the reference standard is imperfect. In general, most ante-mortem reference standards in common use in diagnostic laboratories are imperfect and hence, necropsy with testing of multiple tissues by ancillary tests such as culture and/or histopathology often is necessary if the results of the reference standard are to be considered to be the truth. For most test validation studies for animal diseases, this latter option is not feasible or cost-effective except for a limited number of samples.

1. DSe and DSp with a perfect reference standard

The candidate test may yield results on binary (dichotomous), ordinal (e.g. titre) or continuous scales. For the latter two scales, results need to be dichotomised before DSe and DSp can be calculated, i.e. a cut-off (threshold) needs to be established. Exact binomial 95% CI are recommended for DSe and DSp (Greiner & Gardner, 2000) because the normal approximation may not yield appropriate CI when parameter estimates are close to 1.

Example: indirect enzyme-linked immunosorbent assay (I-ELISA)

		Number of animals	
		Known antibody positive (369)	Known antibody negative (198)
Test results	Positive	287 TP	FP 1
	Negative	FN 82	TN 197
		Diagnostic sensitivity* TP/(TP + FN) **77.8%** (73.2 – 81.9%)*	Diagnostic specificity* TN/(TN + FP) **99.5%** (97.2 – 99.9%)*

TP and FP = true positive and false positive, respectively

TN and FN = true negative and false negative, respectively

*95% exact binomial confidence limits for DSe and DSp

When the reference standard is not applied to all positive and negative test results (partial verification), corrected estimates of DSe and DSp should be made as described in Greiner & Gardner (2000) to account for different sampling probabilities in the test-positive and test-negative groups.

For assays yielding ordinal (e.g. titre values) or continuous results (e.g. ratios of test sample to positive control sample values in an ELISA), estimates of DSe and DSp should be complemented with estimates of the area under the receiver-operating characteristic (ROC) curve. ROC analysis provides a cut-off-independent approach for evaluation of the global accuracy of a test where results are measured as ordinal or continuous values. The area under the ROC curve provides a single numerical estimate of overall accuracy ranging from 0.5 (useless test) to 1 (perfect test). The main justification for ROC analysis is that cut-off values for test interpretation may change depending on the purpose of testing (e.g. screening versus confirmation) and with the prevalence of infection, the costs of test errors, and the availability of other tests. Detailed descriptions of ROC analysis are presented elsewhere (Gardner & Greiner, 2006; Greiner *et al.,* 2000; Zweig & Campbell, 1993). When multiple ordinal or continuous tests are compared, the difference in the area under the curve with a 95% CI should be calculated. Methods for calculating differences vary for independent and dependent samples and are implemented in many statistical programs (Gardner & Greiner, 2006). Examples of a dot diagram for results of a single ELISA and ROC curves for two ELISAs are shown in Figures 5 and 6.

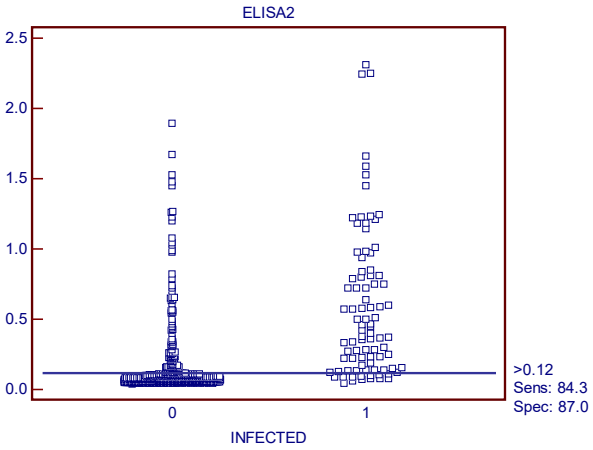


Fig. 5. Dot diagram of ELISA results for non-infected (Code = 0) and infected (Code = 1) animals.

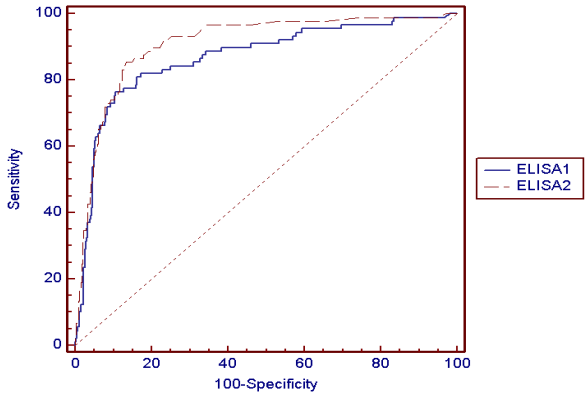


Fig. 6. Receiver-operator characteristic curve for two ELISAs.

In the absence of a perfect reference standard, it is also possible to estimate the AUC using latent class (LC) models. For example, LC models can be applied to normally distributed data from two dependent tests (see for example Choi *et al.,* 2003) and using semiparametric approaches (Branscum *et al.,* 2008). LC models for continuous data including censored or truncated data that occur with real-time PCR assays are not described in this validation guideline because of their complexity. However, LC models for binary test results and an example application are described in Section E.3.

2. Comparison of DSe and DSp estimates for two tests with a perfect reference standard

Often, investigators might wish to compare DSe values in subpopulations of infected animals, e.g. clinically versus subclinically infected, or DSp values in different geographical areas. Since these are independent samples, the comparisons can be made statistically by Pearson's chi-square test for homogeneity. Alternately, separate 95% CI and a 95% CI for a difference in two proportions can be calculated. When the DSe (or DSp) of two tests are compared on the same set of infected (or non-infected) samples in a paired design, the test results are no longer independent. Statistical methods such as McNemar's chi-square can be used to test the hypothesis of equal sensitivities (specificities) when testing is done on the same samples.

Example: Five antibody detection tests were evaluated for diagnosis of bovine paratuberculosis in dairy cows in known infected and non-infected herds, as determined by faecal culture results and herd history. The following data tables were generated based on the original data before subsequent publication in Collins *et al*. (2005). In the publication, one herd was removed from the analysis. The example is used for demonstration purposes to show the tabular layout for calculation of DSe and DSp and statistical evaluation.

Infected

	T_{2+}	T_{2-}	
T_{1+}	124	74	198
T_{1-}	8	243	251
	132	317	449

Non-infected

	T_{2+}	T_{2-}	
T_{1+}	3	27	30
T_{1-}	16	366	382
	19	393	412

Sensitivity of T1 = 198/449 = 44.1%

Sensitivity of T2 =132/449 = 29.4%

Specificity of T1 = 382/412 = 92.7%

Specificity of T2 = 393/412 = 95.4%

Sensitivities differed significantly ($p<0.0001$), but specificities did not ($p= 0.126$) based on a two-tailed McNemar's chi-square test. Sensitivity and specificity covariances (see Gardner *et al*., 2000 for details) can also be calculated to indicate whether the tests are conditionally independent or dependent, given infection status. For these data, the sensitivity covariance (calculated using the infected table on the left) was 0.147 ($p < 0.0001$ by Pearson's chi-square) indicating strong dependence of the two tests when used in infected animals. The specificity covariance (calculated using the non-infected table on the right) was 0.004 ($p = 0.152$ by Pearson's chi-square) indicating no significant dependence.

An additional example based on porcine toxoplasmosis data is presented in Gardner *et al*. (2010).

3. DSe and DSp without a perfect reference standard

Advances in statistical methodology, specifically the development of latent class (sometimes termed "no-gold-standard") models, now allow investigators to liberate themselves from the restrictive assumption of a perfect reference test and estimate the accuracy of the candidate test(s) and the reference standard with the same data (Enoe *et al.,* 2000; Hui & Walter, 1980).

Latent class (LC) models, either using maximum likelihood or Bayesian methods, can be used for estimation of DSe and DSp when joint test results are available from multiple tests applied to animals in multiple populations (e.g. herds or geographical areas). Not all LC models for estimation of DSe and DSp will be statistically identifiable for inference. A model is identifiable if it is theoretically possible to determine the true value of model parameters after obtaining an infinite number of observations from it. In essence, this equates to having a unique set of values for the parameters of interest (DSe, DSp). Bayesian approaches are especially suited to situations where prior information is available about DSe and/or DSp and when the estimation problem is not identifiable (Branscum *et al*., 2005).

The simplest one-population LC model that is identifiable is when three conditionally independent are run on the same samples. The constraint of independence of three tests may be difficult to achieve in practice unless the target analyte differs among tests. Hence, a commonly used approach in animal health is to run two tests on all samples from animals in two populations because it is less costly and assumptions of conditional independence may be more reasonable. The two-test two-population model also requires the assumptions of constant sensitivity and specificity across the two populations, and distinct prevalences. The assumption about constant sensitivity may be difficult to verify and is unlikely to be correct if one population has clinically affected animals and the other population has subclinically affected animals because many published studies have shown that test sensitivity is greater in clinically affected animals. If one of the two populations is known to be pathogen-free (prevalence is zero) while the other population is known to have a non-zero prevalence, the former population can be used for estimation of DSp and this will facilitate estimation of DSe in the infected population.

WOAH-listed diseases where DSe and DSp have been estimated with Bayesian methods include ovine brucellosis (Praud *et al*., 2012), Q fever (Paul *et al*., 2013), trypanosomosis (Bronsvoort *et al*., 2010), bovine tuberculosis (Clegg *et al*., 2011), foot and mouth disease (Bronsvoort *et al*., 2006), African horse sickness (Guthrie *et al*., 2013) and infectious salmon anaemia virus (Caraguel *et al*., 2012).

The WinBUGS software[1] allows easy implementation of Markov-chain Monte Carlo methods for Bayesian estimation (Lunn *et al.,* 2000) and simple maximum likelihood analyses can be done using a web-based interface (Poulliot *et al.,* 2002). Prior information about model parameters used in the Bayesian analyses may affect the final estimates depending on the relative strength of evidence provided by the priors (level of prior uncertainty) and the data (uncertainty attributable to finite sample sizes). Therefore, the sources of prior information must be well documented in Bayesian analyses and it may be desirable to repeat the analysis using non-informative priors on all parameters when the model is identifiable.

Maximum likelihood – a method for estimation of the most likely values for the parameters of interest based on the value of likelihood function for the data.

Bayesian methods – incorporate relevant prior information or knowledge about one or more tests in addition to the likelihood function for the data. With large sample size, maximum likelihood and Bayesian methods will yield similar inferences.

It is important to note that LC analysis cannot correct for biases inherent in poorly designed studies. The methods should be used carefully and include a thorough evaluation of underlying assumptions (e.g. conditional dependence, constant sensitivity and specificity across populations, and distinct prevalences), the effects of use of the selected prior distributions on posterior inferences as described in the previous paragraph, and convergence of Markov chains in a Bayesian analysis (Toft *et al.,* 2005).

Example: Guthrie *et al*. (2013) estimated the DSe and DSp of a quantitative real-time PCR and conventional virus isolation (VI) for detection of African horse sickness (AHS) virus in whole blood samples using a two-test two-population Bayesian latent class model. Two populations of South African thoroughbred horses (503 AHS suspect cases and 503 healthy horses from the AHS virus controlled zone) were tested by PCR and VI. For the 503 suspect cases the joint test results were: PCR+VI+ (*n*=156). PCR+VI– (*n*=184), PCR–VI+ (*n*=0), and PCR–VI– (*n*=163). All 503 healthy horses were PCR–VI–. Various models (conditional independence and conditional dependence) were fitted to the data and a second population of healthy horses was also included in some analyses.

Models were run in WinBUGS 1.4.3 (Lunn *et al.,* 2000) with the first 5000 iterations discarded and the next 50,000 iterations used for posterior inferences (medians and 95% probability intervals for DSe and DSp. Model convergence was assessed by visual inspection of trace plots of iterated values and running multiple chains from dispersed initial values. The conditional independence model fitted with non-informative beta (1,1) priors on DSe and DSp of both tests yielded almost identical results to the model which used a highly informative beta (9999,1) prior for the DSp of VI. Estimated medians values and 95% probability intervals (sometimes termed credible intervals) in parentheses from the conditional independence model with non-informative priors were:

PCR sensitivity = 0.996 (0.977–0.999)

PCR specificity = 0.999 (0.993–1.0)

VI sensitivity = 0.458 (0.404–0.51)

VI specificity = 0.999 (0.998–1.0)

The results indicated a twofold higher DSe of PCR compared with VI and comparable DSp of both tests. For a complete description of the modelling approach see Guthrie *et al.* (2013).

4. Comparison of DSe and DSp estimates for two tests without a perfect reference standard

If a Bayesian approach is used in WinBUGS to analyse the joint test data from multiple populations, the difference in sensitivities (specificities) can be readily estimated and the probability that the sensitivity (specificity) of one test exceeds the other can be estimated with the STEP function.

Example: For the results of the Guthrie *et al.* (2013) data in Section E.3, the 95% probability intervals (PI) for DSe did not overlap but there was marked overlap in the 95% PI for DSp. The corresponding probability values obtained from the STEP function were 1 and 0.24, respectively. These values indicate certainty that the DSe differ but the probability that the DSp differ is small (less than 0.5).

1 Available at http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

REFERENCES

BARTLETT J.W. & FROST C. (2008). Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet. Gynecol.*, **31**, 466–475.

BLAND J.M. & ALTMAN D.G. (1999). Measuring agreement in method comparison studies. *Statist. Methods Med. Res.*, **8**, 135–160.

BOSSUYT P.M., REITSMA J.B., BRUNS D.E., GATSONIS C.A., GLASZIOU P.P., IRWIG L.M., LIJMER J.G., MOHER D., RENNIE D. & H.C.M. DE VET (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin. Chem.,* **49**, 1–6.

BRANSCUM A.J., GARDNER I.A. & JOHNSON W.O. (2005). Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Prev. Vet. Med.*, **68**, 145–163.

BRANSCUM A.J., JOHNSON W.O., HANSON T.E. & GARDNER I.A. (2008). Bayesian semiparametric ROC curve estimation and disease diagnosis. *Stat. Med.*, **17**, 2474–2496.

BRONSVOORT B.M., TOFT N., BERGMANN I.E., SØRENSEN K.J., ANDERSON J., MALIRAT V., TANYA V.N., MORGAN K.L. (2006) Evaluation of three 3ABC ELISAs for foot-and-mouth disease non-structural antibodies using latent class analysis. *BMC Vet. Res.*, **2**, 30.

BRONSVOORT B.M., VON WISSMANN B., FÈVRE E.M., HANDEL I.G., PICOZZI K., & WELBURN S.C. (2010) No gold standard estimation of the sensitivity and specificity of two molecular diagnostic protocols for *Trypanosoma brucei* spp. in Western Kenya. *PLoS One*; **5** (1), e8628.

BURNS M & VALDIVIA H. (2008). Modelling the limit of detection in real-time quantitative PCR. *Eur. Food Res. Technol.*, **226**, 1513–1524.

CARAGUEL C., STRYHN H., GAGNÉ N., DOHOO I. & HAMMELL L. (2012). Use of a third class in latent class modelling for the diagnostic evaluation of five infectious salmon anaemia virus detection tests. *Prev. Vet. Med.,* **104**, 165–173.

CHOI Y.K., JOHNSON W.O., COLLINS M.T. & GARDNER I.A. (2006). Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. *J. Agric, Biol. Environ. Stat.,* **11**, 201–229.

CLEGG T.A., DUIGNAN A., WHELAN C., GORMLEY E., GOOD M., CLARKE J., TOFT N. & MORE S.J. (2011). Using latent class analysis to estimate the test characteristics of the γ-interferon test, the single intradermal comparative tuberculin test and a multiplex immunoassay under Irish conditions. *Vet. Microbiol.*, **151**, 68–76.

COLLINS M.T., WELLS S.J., PETRINI K.R., COLLINS J.E., SCHULTZ R.D., & WHITLOCK R.H. (2005). Evaluation of five antibody detection tests for diagnosis of bovine paratuberculosis. *Clin. Diag. Lab. Immunol.*, **12**, 685-692.

DARGATZ D.A., BYRUM B.A., COLLINS M.T., GOYAL S.M., HIETALA S.K., JACOBSON R.H., KOPRAL C.A., MARTIN B.M., MCCLUSKEY B.J. & TEWARI D. (2004). A multilaboratory evaluation of a commercial enzyme-linked immunosorbent assay test for the detection of antibodies against *Mycobacterium avium* subsp. *paratuberculosis* in cattle. *J. Vet. Diagn. Invest.,* **16**, 509–514.

DONNER A & ZOU G.Y. (2012). Closed-form confidence intervals for functions of the normal mean and standard deviation. *Stat. Meth. Med. Res.*, **21**, 347–359.

ENØE C., GEORGIADIS M.P. & JOHNSON W.O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. Vet. Med.,* **45**, 61–81.

FLEISS J.L., LEVIN B. & PAIK M.C. (2003). Statistical Methods for Rates and Proportions, Third Editionn. John Wiley & Sons, New York, USA.

GARDNER I.A., STRYHN H., LIND P., & COLLINS M.T. (2000). Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Prev. Vet. Med.,* **45**, 107–122.

GARDNER I.A. & GREINER M. (2006). Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. *Vet. Clin. Pathol.*, **35**, 8–17.

GARDNER I.A., GREINER M. & DUBEY J.P. (2010). Statistical evaluation of test accuracy studies for *Toxoplasma gondii* in food animal intermediate hosts. *Zoonoses Public Health*, **57**, 82–94.

GREINER M. & GARDNER I.A. (2000). Epidemiologic issues in the validation of veterinary diagnostic tests. *Prev. Vet. Med.,* **45**, 3–22.

GREINER M., PFEIFFER D. & SMITH R.D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev. Vet. Med.,* **45**, 23–41.

GUTHRIE A.J., MACLACHLAN N.J., JOONE C., LOURENS C.W., WEYER C.T., QUAN M., MONYAI M.S. & GARDNER I.A. (2013). Diagnostic accuracy of a duplex real-time reverse transcription quantitative PCR assay for detect ion of African horse sickness virus. *J. Virol. Methods,* **189**, 30–35.

HUI S.L. & WALTER S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, **36**, 167–171.

LANDIS J.R. & KOCH G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

LUNN D.J., THOMAS A., BEST N. & SPIEGELHALTER D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statist. Comp.,* **10**, 325–337.

PAUL S., TOFT N., AGERHOLM J.S., CHRISTOFFERSEN A.B. & AGGER J.F. (2013). Bayesian estimation of sensitivity and specificity of *Coxiella burnetii* antibody ELISAs in bovine blood and milk. *Prev. Vet. Med.,* **109**, 258–263.

PRAUD A., CHAMPION J.L., CORDE Y., DRAPEAU A., MEYER L. & GARIN-BASTUJI B. (2012) Assessment of the diagnostic sensitivity and specificity of an indirect ELISA kit for the diagnosis of *Brucella ovis* infection in rams. *BMC Vet. Res.,* **8**, 68.

POUILLOT R., GERBIER G. & GARDNER I.A. (2002). "TAGS", a program for the evaluation of test accuracy in the absence of a gold standard. *Prev. Vet. Med.,* **53**, 67–81.

TOFT N., JORGENSEN E. & HOJSGAARD S. (2005). Diagnosing diagnostic tests: evaluating the assumptions underlying the estimated of sensitivity and specificity in the absence of a gold standard. *Prev. Vet. Med.,* **68**, 19–33.

WILKS C. (2001). Gold standards as fool's gold. *Aust. Vet. J.,* **79**, 115.

ZWEIG M.H. & CAMPBELL G. (1993). Receiver-operating characteristic (ROC) plots - a fundamental evaluation tool in clinical medicine. *Clin. Chem.,* **39**, 561–577.

*
* *

NB: FIRST ADOPTED IN 2014.