CHAPTER 1.1.4.

PRINCIPLES OF VALIDATION OF DIAGNOSTIC ASSAYS FOR INFECTIOUS DISEASES

INTRODUCTION

Validation is the evaluation of a process to determine its fitness for a particular use and includes assay optimisation and demonstration of performance characteristics. An assay validated for an infectious disease yields test results that identify the presence of a particular analyte (*e.g. components of an infectious agent or antibody induced by it*) *and allows predictions to be made about the status of the test subjects. Assays applied to individuals or populations have many purposes, such as aiding in: documenting freedom from disease in a country or region, preventing spread of disease through trade, eradicating an infection from a region or country, confirming diagnosis of clinical cases, estimating infection prevalence to facilitate risk analysis, identifying infected animals toward implementation of control measures, and classifying animals for herd health or immune status post-vaccination. A single assay may be validated for one or several intended purposes by optimising its performance characteristics for each purpose* (*e.g. setting diagnostic sensitivity* [*DSe*] *high* [*such as 99.99%*] *with associated lower diagnostic specificity* [*DSp*] *for a screening assay, or conversely, setting DSp high with associated lower DSe for a confirmatory assay*).

The principles of validation discussed in this chapter will focus primarily on methods to detect antibody in sera using an ELISA as an example. However, these same principles are applicable to validation of tests for other analytes in sera or tissues. Chapter 1.1.5 Validation and quality control of polymerase chain reaction methods used for the diagnosis of infectious diseases extends the principles outlined here to a direct method of infectious agent detection, the molecular diagnostic assays.

By considering the variables that affect an assay's performance, the criteria that must be addressed in assay validation become clearer. The variables can be grouped into three categories: (*a*) *the sample – host/organism interactions affecting the analyte composition and concentration in the serum sample;* (*b*) *the assay system – physical, chemical, biological and technician-related factors affecting the capacity of the assay to detect a specific analyte in the sample; and* (*c*) *the test result – the capacity of a test result, derived from the assay system, to predict accurately the status of the individual or population relative to the analyte in question.*

Factors that affect the concentration and composition of analyte in the serum sample are mainly attributable to the host and are either inherent (*e.g. age, sex, breed, nutritional status, pregnancy, immunological responsiveness*) *or acquired* (*e.g. passively acquired antibody, active immunity elicited by vaccination or infection*). *Nonhost factors, such as contamination or deterioration of the sample, may also affect the analyte in the sample.*

Factors that interfere with the analytical accuracy of the assay system include instrumentation, technician error, reagent choice (*both chemical and biological*) *and calibration, accuracy and acceptance limits of controls, reaction vessels, water quality, pH and ionicity of buffers and diluents, incubation temperatures and durations, and error introduced by detection of closely related analytes, such as antibody to cross-reactive organisms, rheumatoid factor, or heterophile antibody.*

Measures that influence the capacity of the test result to predict accurately the infection or analyte status of the host[1] *are DSe, DSp, and prevalence of the disease in the population targeted by the*

1 In this chapter, the terms 'positive' and 'negative' have been reserved for test results and never refer to infection or antibody/antigen status of the host. Whenever reference is made to 'infection' or 'analyte', any method of exposure to an infectious agent that could be detected directly (e.g. antigen) or indirectly (e.g. antibody) by an assay, should be inferred.

assay. DSe and DSp are derived from test results on samples obtained from selected reference animals. The methods used to select the reference animals are critical to the accuracy of the estimates (*5*). *The degree to which the reference animals represent all of the host and environmental variables in the population targeted by the assay has a major impact on the accuracy of test-result interpretation. For example, experienced diagnosticians are aware that an assay, validated by using samples from northern European cattle, may not give valid results for the distinctive populations of cattle in Africa.*

The capacity of a positive or negative test result to predict accurately the infection and/or exposure status of the animal or population of animals is the most important consideration of assay validation. This capacity is not only dependent on a highly precise and accurate assay (incorporating well-characterised and standardised reagents) and carefully derived estimates of DSe and DSp, but is heavily influenced by prevalence of the infection in the targeted population or the likelihood that an animal is infected based on clinical criteria. Without a current estimate of the disease prevalence in that population or likelihood of infection in an individual animal, the interpretation of a positive or negative test result may be compromised.

Many factors obviously must be addressed before an assay can be considered to be 'validated' (*5, 16*). *However, there is no consensus whether the concept of assay validation is a time-limited process during which only those factors intrinsic to the assay are optimised and standardised, or whether the concept includes an ongoing validation of assay performance for as long as the assay is used. Accordingly, the term 'validated assay' elicits various interpretations among laboratory diagnosticians and veterinary clinicians. Therefore, a working definition of assay validation is offered as a context for the guidelines outlined below. Ideally, all diagnostic assays would be fully validated for one or more purposes, but in practice there are sometimes limitations to the completeness of validation.*

DEFINITIONS OF ASSAY VALIDATION

A validated assay consistently provides test results that identify animals as positive or negative for an analyte or process (e.g. antibody, antigen, or induration at skin test site) and, by inference, accurately predicts the infection and/or exposure status of animals with a predetermined degree of statistical certainty[2]. Implicit in this definition is the requirement that the test method was properly developed, optimised, and standardised to achieve performance characteristics that are consistent with the purpose for which the assay is intended.

This chapter will focus on the principles underlying development and maintenance of a validated assay. Previous iterations of this chapter (12) were condensed renditions of a review article (9). At that time, the goal was to provide general principles of assay validation. In this update, the content is reorganised into the parts of assay validation consistent with the format of the OIE Validation Template, and emphasises the absolute necessity of pre-determining the specific purpose(s) for which the assay is intended. In addition to the validation process *per* se, guidance is offered on scientifically sound methods for development, maintenance, and extension of validation criteria for a given assay.

It must be emphasised that an assay, when applied to target populations, will minimise misclassifications of animals as false positive or false negative only to the extent that validity is assured for all elements of the assay validation process (see Section B. Assay Validation – Part I). This assumes that the assay is fit for the purpose for which it is intended (e.g. a confirmatory assay will likely yield many false-negative results if used as a screening assay). It also assumes that a well conceived, designed, and documented test method and proper standardised reagents, in combination with well-trained technicians, will give a stable assay within the laboratory. Furthermore, it assumes a thorough use of the most rigorous experimental design and epidemiological and statistical tools. These are required to reduce bias, random error, and false assumptions about the reference population of

2 In this definition, the DSe and DSp are performance characteristics of the assay for a given target population. They determine – together with the disease prevalence in the population – the probability that a given test result reflects the true status of the animal. An assay can be recognised as validated if reliable estimates of DSe and DSp for a given target population are available. This does not imply any minimum threshold values for these parameters. In practical applications, low values of DSe and DSp or diagnostic problems due to low disease prevalence are compensated by the sampling design or by combining multiple diagnostic assays into parallel or serial testing regimens. The selection of assays, the sampling process, the combination of multiple assays into a testing regimen and the interpretation rule for the results define the diagnostic process.

animals upon which the assay performance estimates are made (5). Finally, it assumes that when placed in practice, the assay is conducted within the context of a rigorous quality assurance programme.

A. ESSENTIAL PREREQUESITES BEFORE VALIDATION OF AN ASSAY

1. Selection of an assay fit for its intended purpose

The OIE *Standard for Management and Technical Requirements for Laboratories Conducting Tests for Infectious Diseases* (14)[3]. This Standard states that test methods and related procedures must be appropriate for specific diagnostic applications in order for the test results to be of any relevance. In other words, the assay must be 'fit for purpose'.

As outlined in the background information in *Certification of diagnostic assays* on the OIE website (www.oie.int), the first step is selection of an assay type that likely can be validated for a particular use. The intended purpose(s) of an assay have been broadly defined:

1) Demonstrate freedom from infection in a defined population (country/zone/compartment/herd) (prevalence apparently zero):
   - 1a) 'Free' with and/or without vaccination,
   - 1b) Historical freedom,
   - 1c) Re-establishment of freedom after outbreaks.
2) Certify freedom from infection or agent in individual animals or products for trade/movement purposes.
3) Eradication of infection from defined populations.
4) Confirmatory diagnosis of suspect or clinical cases (includes confirmation of positive screening test).
5) Estimate prevalence of infection or exposure to facilitate risk analysis (surveys, herd health status, disease control measures).
6) Determine immune status of individual animals or populations (post-vaccination).

The OIE Standard further states that in order for a test method to be considered appropriate, it must be properly validated and that this validation must respect the principles outlined in the validation chapters of the this *Terrestrial Manual*.

While this chapter deals with validation and fitness for purpose from a scientific perspective, it should also be noted that other factors might impact the relevance of an assay with respect to fitness for purpose. These factors include not only the diagnostic suitability of the assay, but also its acceptability by scientific and regulatory communities, acceptability to the client, and feasibility given available laboratory resources. An inability to meet operational requirements of an assay also may make it unfit for its intended purpose. Such requirements may include running costs, equipment availability, level of technical sophistication and interpretation skills, kit/reagent availability, shelf life, transport requirements, safety, biosecurity, sample throughput, test turn-around times, aspects of quality control and quality assurance.

2. Initial assay development considerations

An indirect enzyme-linked immunosorbent assay (ELISA) for detection of antibody will be used in this chapter to illustrate the principles of assay validation. It is a type of assay that can be difficult to validate because of signal amplification of both specific and nonspecific components (2). This methodology serves to highlight the problems that need to be addressed in any assay validation process. The same basic principles are used in validation of other complex or simple assay formats. However, each unique type of assay, such as an antigen detection ELISA, may have different sample collection and storage requirements. This chapter assumes that the assay developer will have a high level of relevant scientific expertise to achieve proper preparation and use of protocols and reagents, leading to a validated assay that is publishable in peer reviewed journals. Chapter 1.1.5 Validation and quality control of polymerase chain reaction methods used for the diagnosis of infectious diseases describes the principles for validating gene-amplification techniques.

Selection of appropriate samples, calibrated instrumentation and a relevant methodology to achieve the intended purpose are critical elements in assay validation. Continuity in experiments is assured when reagents and samples are chosen, properly prepared, aliquoted, and stored for use in each experiment. This reduces to a minimum the number of variables and guards against failure when the validation process commences. This

3 This is a specific interpretation of the more generally stated requirements of the ISO/IEC 17025:2005 international quality standard for testing laboratories (8).

approach reduces the variability and provides data needed to establish appropriate controls for ensuring each run of the assay is valid.

a) Feasibility studies - selection of samples

Samples are needed for experiments to determine if the proposed assay is feasible. For this preliminary step, it is useful to select four or five sera (in our example) that range from high to low levels of antibodies against the infectious agent in question. In addition, a sample containing no antibody is required. These samples ideally should represent known infected and uninfected animals from the population that eventually will become the target of the assay once it is validated. The samples should have given expected results in one or more serological assay(s) other than the one being validated. The samples are preferably derived from individual animals, but they may represent pools of samples from several animals. These samples can be used in experiments to determine if the assay is able to distinguish between varying quantities of analyte (antibody in our example), and for optimising the reagent concentrations and perfecting the protocol.

A good practice is to prepare a large volume (e.g. 10 ml or more if possible) of each sample and divide it into 0.1 ml aliquots for storage at or below –20°C. One aliquot of each sample is thawed, used for experiments, and ideally then discarded. If it is impractical to discard the aliquot, it may be held at +4°C between experiments for up to about 2 weeks; however, there is a possibility of sample deterioration under these circumstances. Then, another aliquot is thawed for further experimentation. This method provides the same source of serum with the same number of freeze–thaw cycles for all experiments (repeated freezing and thawing of serum can denature antibodies so should be avoided). Also, variation between experiments is reduced when the same source of serum is used for all experiments rather than switching among various sera between experiments. This approach has the added advantage of generating a data trail for the repeatedly run samples.

Repeated runs using these samples also can provide preliminary repeatability assessments both within and between runs of the assay. When compared with international standards to establish their activity (concentration or titre), one or more of these samples may also serve as secondary standards; such standards provide assurance that runs of the assay are producing accurate data (16).

Finally, these pools of sera may be used as controls in future routine runs of the assay once all steps of the validation process have been completed.

It is highly desirable to include OIE International Standard Sera or other international standard sera (if they are available) at an early stage in assay development. This may facilitate harmonisation between the assay under development and a standard test method in which international standard sera are normally used (15).

b) Selection of method to achieve normalised results

Normalisation adjusts raw test results of all samples relative to values of controls included in each run of the assay (not to be confused with transformation of data to achieve a 'normal' [Gaussian] distribution). The method of normalisation and expression of data should be determined, preferably no later than at the end of the feasibility studies. Comparisons of results from day to day and between laboratories are most accurate when normalised data are used. For example, in ELISA systems, raw optical density (absorbance) values are absolute measurements that are influenced by ambient temperatures, test parameters, and photometric instrumentation. To account for this variability, results are expressed as a function of the reactivity of one or more serum control samples that are included in each run of the assay. Data normalisation is accomplished in the indirect ELISA by expressing absorbance values in one of several ways (16). A simple and useful method is to express all absorbance values as a percentage of a single high-positive serum control that is included in each plate (Sample/Positive or S/P ratio). (This control must yield a result that is in the linear range of measurement.) This method is adequate for most applications. More rigour can be brought to the normalisation procedure by calculating results from a standard curve generated by several serum controls. It requires a more sophisticated algorithm, such as linear regression or log-logit analysis. This approach is more precise because it does not rely on only one high-positive control sample for data normalisation, but rather uses several serum controls, adjusted to expected values, to plot a standard curve from which the sample value is extrapolated. This method also allows for exclusion of a control value that may fall outside expected confidence limits.

For assays that are end-pointed by sample titration, such as serum (viral) neutralisation, each run of the assay is accepted or rejected based on whether control values fall within predetermined limits. Because sample values usually are not adjusted to a control value, the data are not normalised by the strict definition of the term.

Whatever method is used for normalisation of the data, it is essential to include additional controls for any reagent that may introduce variability and thus undermine attempts to achieve a validated assay. The

normalised values for those controls need to fall within predetermined limits (e.g. within an appropriate multiple of the standard deviation of the mean of many runs of each control). The chosen limits should reflect a reasonable and tolerable assay run rejection rate and an acceptable risk that some test samples may be misclassified.

B. ASSAY VALIDATION - PART 1

1. Optimisation and standardisation of reagents

Using several well-defined sera, such in-house standards as outlined in Section A.2.a of this chapter, or reference standards from outside sources, the optimal concentrations/dilutions of the antigen adsorbed to the plate, serum, enzyme–antibody conjugate, and substrate solution are determined through 'checkerboard' titrations of each reagent against all other reagents, following confirmation of the best choice of reaction vessels (usually evaluation of two or three types of microtitre plates, each with its different binding characteristics, to minimise background activity while achieving the maximum spread in activity between negative and high-positive samples). Additional experiments determine the optimal temporal, chemical, and physical variables in the protocol, including incubation temperatures and durations; the type, pH, and molarity of diluent, washing and blocking buffers; and equipment used in each step of the assay (for instance pipettes and washers that give the best reproducibility).

The choice of reagents and their characterisation must be carefully addressed, or the assay's performance characteristics likely will be compromised. For example, increased assay specificity can be accomplished through recombinant expression of antigens or by use of monoclonal antibodies in antigen capture or antibody competition assays. Alternatively, the method of reagent production can also lead to reduced specificity and increased variability. For example, if a viral antigen used in the assay is derived from a viral culture system that is also used to produce viral vaccines commonly used in the species targeted by the assay, nonspecific cross reactivity may occur. Absorption of cross reactive antigens that are in both the vaccine and the antigen used in the assay is necessary, or a cell culture control needs to be tested on each serum sample in routine runs of the assay to identify and account for the extent of such cross reactivity. Obviously, anticipation of the negative or positive impacts of reagent choice on the assay under development/validation is a major consideration, and careful experimentation is necessary to establish an optimal assay.

When a reagent such as a serum control sample is nearing depletion, it is essential to prepare and repeatedly test a replacement before such a control is depleted. The prospective control sample is included in 10–20 runs of the assay before depletion of the original control to establish its proportional relationship to the nearly depleted control. If the depleted sample was a positive control in ELISAs where the normalised value is expressed as a per cent of that positive control, the proportional difference in ELISA activity between the original and replacement sera must be factored into the normalisation algorithm to retain the same cut-off, and thus the same DSe and DSp in the assay. When other reagents, such as antigen for capture of antibody, must be replaced, they should be produced using the same criteria as for the original reagents, and tested in at least five runs of the assay using a panel of sera that has been designed for this purpose. Reagent lots (serials) need to be evaluated for consistency so variability is minimised in the assay as new lots are required. Whenever possible, it is important to change only one reagent at a time to avoid the compound problem of evaluating more than one variable at a time. Variability is minimised when reagents are well-characterised using methods other than that of the target assay.

a) Linear operating range of the assay

The range of values that constitute the linear operating range of an assay is best determined by a dilution series in which a high positive serum is serially diluted in a negative serum. Each dilution is then run at the optimal working dilution in buffer, and the results plotted in the form of a 'response-curve'. This curve, sometimes referred to as a 'dose–response curve' as in pharmacological applications, establishes the linear range of assay values that are valid for use in the assay.

b) Calibration against reference reagents

i) *International standards*

Serum standards and other reagents, available from OIE, WHO, FAO, or other international organisations, can be used to harmonise the assay with expected results gained from reference reagents of known activity.

ii) *In-house standards*

The in-house serum controls (used for normalisation of data) and additional secondary serum standards, such as low positive, high positive, and negative sera (used for repeatability estimates in

subsequent routine runs of the assay) can be fitted to the response curve to achieve expected values for such sera.

2. Repeatability

Preliminary evidence of repeatability (agreement between replicates within and between runs of the assay) is necessary to warrant further development of the assay. This is accomplished by evaluating results from a minimum of three in-house samples representing activity within the linear range of the assay. Quadruplicates of these samples are tested in at least four runs of the assay to determine within-run (intraplate) variation. Between-run variation is determined by using the same samples in a minimum of 20 runs (total), by two or more operators, preferably on separate days. All runs must be independent of each other.

For reporting purposes, ELISA raw absorbance values are usually used to calculate repeatability during this part of validation because it is uncertain whether the results of the high-positive control serum, which could be used for calculating normalised values, are reproducible in early runs of the assay format. Also, expected values for the controls have not yet been established. Coefficients of variation (CV: standard deviation of replicates ÷ mean of replicates), generally less than 20% for raw absorbance values for most samples (low-titred samples may have larger CVs), indicates adequate repeatability at this stage of assay development. However, if evidence of excessive variation (>30%) is apparent for most samples within and/or between runs of the assay, more preliminary studies should be done to determine whether stabilisation of the assay is possible, or whether the test format should be abandoned. This is important because an assay that is inherently variable has a high probability of not withstanding the rigours of day-to-day testing on samples from the targeted population of animals.

Additional evidence of repeatability is obtained from the many additional runs of the assay that are required later in the validation process to fully validate the assay. This is accomplished by running replicates of each control, standard, and test sample when experiments are conducted to establish other validation parameters (see Section C. Assay Validation – Part 2, below). Such data will lend confidence to repeatability estimates because they will be based on within-run and between-run assay performance using reagents prepared daily, including different lots of reagents that could affect repeatability.

3. Determination of analytical specificity and sensitivity

Analytical specificity of the assay is the degree to which the assay does not cross-react with other analytes and analytical sensitivity is the smallest detectable amount of the analyte in question, i.e., the lowest detection limit of the assay.

Analytical specificity is assessed by use of a panel of samples derived from animals that have been exposed to genetically related organisms that may stimulate cross-reactive antibodies, or sera from animals with similar clinical presentations. This 'near neighbour analysis' is useful in determining the probability of false-positive reactions in the assay. It is also appropriate to document a group specificity criterion that includes detection of the analyte of interest in sera from animals that have experienced infections/exposure to an entire group or serotype of organisms of interest. It is also important to evaluate the analytical specificity of the assay using samples from animals that have been vaccinated. If the assay targets antibody elicited by a virus, vaccination against that virus may produce antibody that interferes with the assay's inferences about infection. Also, if the viral antigen used in the assay is derived from a whole-cell viral culture preparation, containing antigenic reagents (carrier proteins, etc.) in addition to the virus, a vaccinated animal may test falsely positive due to detection of nonviral antibodies.

Analytical sensitivity of an assay can be assessed by quantifying the least amount of analyte that is detectable in the sample. This can be done by limiting dilutions of a standard of known concentration of the analyte. However, such an objective absolute measure is often impossible to achieve due to lack of samples or standards of known concentration or activity. Another approach is to use end-point dilution analysis of samples from known positive animals, to define the penultimate dilution of sample in which the analyte is no longer detectable, or at least, is indistinguishable from the activity of negative sera. When the results for the assay under development are compared with other assay(s) run on the same samples, a relative measure of analytical sensitivity can be estimated.

In addition to analyte standards or samples for which titers have been determined by other assays, it is possible to create samples by spiking a negative sample matrix with known amounts of the analyte in question. In this case, however, spiked samples may be intrinsically different from samples obtained from clinical cases, thus leading to inferences that may not be accurate.

If the intended purpose of the assay is for screening of animals for antibody activity, analytical sensitivity needs to be high to achieve the greatest probability possible for detecting infected animals. If very high analytical sensitivity is not achievable, the assay may not be fit as a screening assay. Alternatively, if confirmation of another independent diagnostic procedure is the purpose for which the assay is intended, analytical specificity is required

that minimises the amount of cross-reactivity. If neither of these objectives is obtainable, the reagents need to be recalibrated, replaced, or the assay should be abandoned.

C. ASSAY VALIDATION - PART 2

1. Determining assay performance characteristics after establishment of a standard assay method and reagent criteria

Estimates of DSe and DSp are the primary performance indicators established during validation of an assay. These must be established after the assay and reagents are optimised and standardised; alteration of protocols or reagents may require reestablishment of performance characteristics. They are the basis for calculation of other parameters from which inferences are made about test results. Therefore, it is imperative that estimates of DSe and DSp are as accurate as possible. Ideally, they are derived from testing a series of samples from reference animals of known history and infection status relative to the disease/infection in question and relevant to the country or region in which the test is to be used, but that is not always possible. A sampling design must be chosen that will allow estimation of diagnostic performance characteristics. However this is a difficult process complicated by logistical and financial limitations. It is also limited by the fact that reference populations and gold standards may be lacking. The following are examples of reference populations and methodologies that may aid in determining performance characteristics of the test being validated.

a) Reference animal populations

i) *Infected or exposed and uninfected or nonexposed reference animals*

Selection of reference animals to evaluate performance characteristics requires that the variables attributable to the target population are represented in the infected/exposed and uninfected/unexposed reference animal populations. The variables include but are not limited to species, age, sex, breed, nutritional status, pregnancy, stage of infection, immunological status including vaccination history, and historical, epidemiological, and/or clinical data including herd disease history should be noted and considered.

ii) *Reference animal status determined by other assays*

In serology, the ‘standard of comparison’ is the results of a method or combination of methods with which the new assay is compared. Although the term ‘gold standard’ is commonly used to describe any standard of comparison, it should be limited to methods that unequivocally classify animals as infected/exposed or uninfected. Some isolation methods themselves have problems of repeatability and sensitivity. Gold standard methods include unequivocal isolation of the agent or pathognomonic histopathological criteria.

Because a true gold standard may be lacking or is impossible to achieve, relative standards of comparison are often necessary; the most common of these include results from other serological assays. Calculations of DSe and DSp are most reliable when the gold standard of comparison is available. When only relative standards of comparison are available, estimates of DSe and DSp for the new assay may be compromised because the error in the estimates of DSe and DSp for the relative standard is carried over into those estimates for the new assay. Indeed, when using imperfect reference tests without efforts to control for any biases, the DSe and DSp performance estimates of the new test will be flawed and thus unacceptable.

iii) *Experimentally infected or vaccinated reference animals*

Sera obtained sequentially from experimentally infected or vaccinated animals have been used to ‘validate’ a new assay. Such repeated observations, pre- and post-seroconversion, from the same animals are not acceptable for establishing estimates of DSe and DSp because the statistical requirement of independent observations is violated. Thus, time-point sampling of individual experimental animals is necessary. Also, exposure to organisms under experimental conditions, or vaccination may elicit antibody responses that are not quantitatively and qualitatively typical of natural infection in the target population (9). The strain of organism, dose, and route of administration to experimental animals are examples of variables that may introduce error when extrapolating DSe and DSp estimates to the target population. For these reasons, validation of an assay should not be based solely on experimental animals.

iv) *Reference animals – Status unknown*

When it is not possible to assemble sera from animals of known infection status, it is possible to estimate DSe and DSp by non-gold standard methods or latent class models (3, 7). Because these statistical models are complex, an expert should be consulted to provide assistance on proper ways to

conduct and describe the sampling from the target population(s), the characteristics of other tests included in the analysis, the appropriate choice of model and the estimation methods based on peer-reviewed literature.

2. Threshold determination

To achieve performance estimates of DSe and DSp of the new assay, the test results first must be reduced to categorical (positive or negative) status. This is accomplished by insertion of a cut-off point (threshold or decision limit) on the continuous scale of test results. Although many methods have been described for this purpose, three examples will illustrate different approaches, together with their advantages and disadvantages. The first is a cut-off based on the frequency distributions (9) of test results from uninfected and infected reference animals. This cut-off can be established empirically by visual inspection of the frequency distributions, by receiver-operator characteristics (ROC) analysis (6, 17), or by selection that favours either DSe or DSp, depending on the intended use for a given assay (11). A second approach is establishing a cut-off based only on uninfected reference animals, for example the 99th percentile in a frequency distribution of assay values for uninfected reference animals; this provides an estimate of DSp but not DSe. The third method provides an 'intrinsic cut-off' based on test results from sera drawn randomly from within the target population with no prior knowledge of the animals' infection status (4).

If considerable overlap occurs in the distributions of test values from known infected and uninfected animals, it is difficult to select a cut-off that will accurately classify these animals according to their infection status. Rather than a single cut-off, two cut-offs can be selected that define a high DSe (e.g. inclusion of 99% of the values from infected animals), and a high DSp (e.g. 99% of the values from uninfected animals). The values that fall between these percentiles would then be classified as suspicious or equivocal, and would require testing by a confirmatory assay or retesting for detection of seroconversion.

The selection of the cut-off will typically reflect the intended purpose of the assay. For example, a screening assay designed for high DSe versus a confirmatory assay designed for high DSp will require different cut-offs in the same assay system. Although the intended purpose will dictate the cut-off, a ROC analysis is still desirable, as it will show the potential performance of the assay in other epidemiological settings.

3. Assay performance estimates

a) Number of reference animals required

The number and source of reference samples coupled with the methodologies used to derive DSe and DSp estimates are of paramount importance if the assay is ever to be properly validated for use in the population of animals targeted by the assay. It is possible to calculate the number of reference samples, from animals of known infection/exposure status, required for determinations of DSe and DSp that will have statistically defined limits. Formulae and tables for determining the number of samples required are provided elsewhere (5, 9). Table 1, page 474 of reference 9 reveals how the number of samples tested affects the confidence levels in the calculated estimates of DSe and DSp for the assay. For example, an estimated DSe or DSp of 92% with a confidence level of 75% in that estimate requires 161 analyte-positive (known infected) animals from the population targeted by the assay (with an allowable error of ± 2%). However, to increase confidence in the estimate to a 95% level requires that 542 samples/animals be tested. The number of samples theoretically required to achieve confidence levels ranging from 75% to 99% can be found in this reference table for assays that are anticipated to have DSe or DSp ranging from 80% to 99%.

b) DSe and DSp estimates based on reference animals with defined infection status

The selection of a cut-off allows classification of test results into positive or negative categories. Calculations of DSe and DSp are aided by associating the positive/negative categorical data with the known infection status for each animal using a two-way (2 × 2) table (Table 1). After the cut-off is established, results of tests on standard sera can be classified as true positive (TP) or true negative (TN) if they are in agreement with those of the gold standard (or other standard of comparison). Alternatively, they are classified as false positive (FP) or false negative (FN) if they disagree with the standard. Diagnostic sensitivity is calculated as TP/ (TP + FN) whereas diagnostic specificity is TN/ (TN + FP); the results of both calculations are usually expressed as percentages (Table 1).

Table 1. *Calculations of DSe and DSp aided by a 2 × 2 table that associates infection status with test results from 2000 reference animals*

		Reference animals of known infection status	
		Infected (*n* = 600)	**Uninfected (*n* = 1400)**
Test Result	**Positive**	**570** **TP**	**46** **FP**
	Negative	**30** **FN**	**1354** **TN**
		Diagnostic sensitivity $\frac{TP}{TP+FN} = \frac{570}{600} = 95.0\%$	**Diagnostic specificity** $\frac{TN}{TN+FP} = \frac{1354}{1400} = 96.7\%$

c) DSe and DSp estimates based on animals with infection status not defined

DSe and DSp can be estimated when infection or analyte status of the animals are not defined; however, these latent class statistical models are complex. Expert advice should be sought not only in the design of the evaluation study but the interpretation of the estimates of DSe and DSp as well. It has been recommended to the OIE that an expert group be formed to address the application of latent class models and to draft guidelines for models as they apply to the validation and certification assays by the OIE.

4. Comparison and harmonisation of assays

New assays usually are developed to improve on existing techniques. In order to demonstrate that a new assay is an improvement over an existing technique, there must be some form of comparison that demonstrates the improvement. The comparison may be related to analytical and/or diagnostic performance characteristics. It may also be related to operational characteristics such as cost, ruggedness, turn-around-times, throughput, etc. If the new assay is to be incorporated into a diagnostic regimen involving other test methods, the rationale for its use, interpretation of data and decision-making should be stated.

When an international standard method (15) is available for detection of an analyte, it is possible to harmonise the performance of that method with the one under development. This process requires use of the same serum controls and/or standards in both assays. If OIE Standard Sera or other international standard sera are available, preferably at least three (negative, low positive, and high positive), they should be included in the assay-comparison study. This could lead to a new assay that is indexed to an international standard method and international standard sera (15). Harmonisation of the two assays may then be realised.

It is critical that all samples, test reagents, and the protocol or instructions for running the assay be properly controlled. If the reagents will not be supplied from a common source, the laboratories should produce and characterise the reagents independently. This will allow determination of the adequacy of the protocol for reagent production and characterisation. This provides data needed to determine whether it is necessary to establish a single shared source of well-characterised reagents. Part of the evaluation is the determination that the protocol or instructions are complete, clear and precise. If verbal instructions are required, the developer should consider revision of the protocol to ensure they are comprehensive. If it is determined that the protocol or instructions were interpreted in a different manner, then they should be rewritten and the reproducibility may need to be re-established using the revised protocol or instructions.

D. ASSAY VALIDATION - PART 3

1. Establishing reproducibility and augmenting repeatability estimates of the assay

An assay intended for distribution to many laboratories (such as a commercial kit) must be evaluated for reproducibility, which is defined as the ability of a test method to provide consistent results when applied to aliquots of the same samples tested at different laboratories. This is accomplished by testing a panel of sera in a minimum of three laboratories using the identical test method and serum panels.

A test panel consisting of a minimum of 20 samples is assembled for this purpose. Ideally, these will be individual samples from animals within the target population, representing the range of assay activity anticipated in that population. If such samples are not available, dilution of a high positive with a negative serum to achieve the range of activity is acceptable but not optimal. Replicates of about 20% of the samples are desirable as a check on repeatability within each participating laboratory. Each sample is aliquoted, rendering a series of identical panels for distribution to other laboratories. The sample identity is encoded for blind testing, and each panel is handled, transported to participating laboratories, and stored identically.

The descriptive statistics for test panel data accumulated from the laboratories includes between-laboratory mean, standard deviation, and range of results for each sample as well as controls. Evaluation of precision and accuracy at each laboratory is facilitated by Youden plots. The data will help to inform the legitimacy of the upper and lower control limits of the assay as established by the developer.

In addition, when the panel of samples is tested in each laboratory, it is advisable to run each sample in duplicate or triplicate. This provides a basis for an expanded analysis of repeatability within each laboratory using the assay. Also, when the assay is placed into routine use, repeatability is also monitored by inclusion of at least duplicates for the controls and preferably for each sample as well.

E. ASSAY VALIDATION - PART 4

1. Programme implementation

Ultimate proof of the usefulness of an assay is its successful application(s). These would include international, regional or national programs. As new and improved assays are developed and come on-line, they will ultimately replace existing assays if they prove a better fitness for purpose. However, this will only happen if they are actually put into routine use and their usefulness documented over time. In the natural progression of diagnostic and/or technological improvement, some new assays will become the new standard of comparison. As such, they may progressively achieve national, regional and international recognition. As a recognised standard, these assays will also be used to develop reference reagents for quality control, proficiency and harmonisation purposes. These reference reagents may also become international standards, as well. The last level of validation in the OIE Registry involves documentation related to actual application and levels of recognition for the assay in question. This is intended to provide potential users with an informed and unbiased source of information.

2. Monitoring validity of assay performance

a) Interpretation of test results - factors affecting assay validity

An assay's test results are useful only if the inferences made from them are accurate. A common error is to assume that an assay with 99% DSe and 99% DSp will generate one false-positive and one false-negative result for approximately every 100 tests on animals from the target population. Such an assay may be precise and accurate, but produce test results that do not accurately predict infection status. For example, if the prevalence of disease in a population targeted by the assay is only 1 per 1000 animals, and the false-positive test rate is 1 per 100 animals (99% DSp), for every 1000 tests on that population, ten will be false positive and one will be true positive. Hence, only approximately 9% of positive test results will accurately predict the infection status of the animal; the positive test results will misclassify the animal 91% of the time. This illustrates that the capacity of a positive or negative test result to predict infection status is dependent on the prevalence of the infection in the target population (10). Of course, the prevalence will probably have been determined by use of a serological test with its own inherent misclassification of results.

An estimate of prevalence in the target population is necessary for calculation of the predictive values of positive (PV+) or negative (PV–) test results. When test values are reported without providing estimates of the assay's DSp and DSe, it is not possible to make informed predictions of infection status from test results (9). It is, therefore, highly desirable to provide an interpretation statement with test results accompanied by a small table indicating PV+ and PV– for a range of expected prevalences of infection in the target population. Without provision of such information, test results from the assay may have failed to accurately classify the infection status of animals, and thus do not reflect a fully validated assay.

b) Maintenance of validation criteria

A validated assay needs constant monitoring and maintenance to retain that designation. Once the assay is put into routine use, internal quality control is accomplished by consistently monitoring the assay for assessment of precision and accuracy (1).

Reproducibility between laboratories should be assessed at least twice each year. It is highly desirable to become part of a consortium of laboratories that are interested in evaluating their output. In the near future,

good laboratory practice, including implementation of a total quality assurance programme, will become essential for laboratories seeking to meet national and international certification requirements (see Chapter 1.1.3 Quality management in veterinary testing laboratories).

Proficiency testing is a form of external quality control for an assay. It is usually administered by a reference laboratory that distributes panels of samples, receives the results from the laboratories, analyses the data, and reports the results back to the laboratories. If results from an assay at a given laboratory remain within acceptable limits and show evidence of accuracy and reproducibility, the laboratory may be certified by government agencies or reference laboratories as an official laboratory for that assay (13). Panels of sera for proficiency testing should contain a full representation of an analyte's concentration in animals of the target population. If the panels only have high-positive and low-positive sera (with none near the assay's cut-off), the exercise will only give evidence of reproducibility at the extremes of analyte concentration, and will not clarify whether routine test results on the target population properly classify infection status of animals.

c) Enhancement and extension of validation criteria

Because of the extraordinary set of variables that impact on the performance of serodiagnostic assays, it is highly desirable to expand the number of standard sera from animals of known infection status because of the principle that confidence in the estimates of DSe and DSp is enhanced with increasing sample size. Furthermore, when the assay is to be applied in a completely different geographical region, it is essential to re-validate the assay for its new intended use by subjecting it to sera from populations of animals that reside under local conditions. The same is true for establishing DSe and DSp for subpopulations (e.g. age groups, vaccinated/nonvaccinated, etc.).

REFERENCES

1. CEMBROWSKI G.S. & SULLIVAN A.M. (1992). Quality control and statistics. *In:* An Introduction to Clinical Chemistry, Bishop M.L., Duben-Engelkirk J.L. & Fody E.P., eds. Lippincott, Philadelphia, USA, 63–101.

2. CROWTHER J.R. (1995). ELISA theory and practice. *In:* Methods in Molecular Biology. Humana Press, Totowa, NJ, USA, 1–256.

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

4. GREINER M., FRANKE C.R., BOHNING D. & SCHLATTMANN P. (1994). Construction of an intrinsic cut-off value for the sero-epidemiological study of *Trypanosoma evansi* infections in a canine population in Brazil: a new approach towards unbiased estimation of prevalence. *Acta Trop.*, **56**, 97–109.

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

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

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

8. INTERNATIONAL ORGANIZATION FOR STANDARDIZATION/INTERNATIONAL ELECTROTECHNICAL COMMISSION (ISO/IEC) (2005). ISO/IEC 17025:2005, General requirements for the competence of testing and calibration laboratories.

9. JACOBSON R.H. (1998). Validation of serological assays for diagnosis of infectious diseases. *Rev. sci. tech. Off. int. Epiz.*, **17,** 469–486.

10. JACOBSON R.H. (1991). How well do serodiagnostic tests predict the infection of disease status of cats. *J. Am. Vet. Med. Assoc.*, **199**, 1343–1347.

11. SMITH R.D. (1991). Clinical Veterinary Epidemiology. Butterworth-Heinemann, Stoneham, MA, USA, 1–223.

12. WORLD ORGANISATION FOR ANIMAL HEALTH (OIE) (1996). Principles of validation of diagnostic assays for infectious diseases. *In:* OIE Manual of Standards for Diagnostic Tests and Vaccines, Third Edition. OIE, Paris, France, 8–15.

13. WORLD ORGANISATION FOR ANIMAL HEALTH (OIE) (2002). OIE Guide 3: Laboratory Proficiency Testing. *In:* OIE Quality Standard and Guidelines for Veterinary Laboratories: Infectious Diseases. OIE, Paris, France, 53–63.

14. WORLD ORGANISATION FOR ANIMAL HEALTH (OIE) (2002). OIE Standard for Management and Technical Requirements for Laboratories Conducting Tests for Infectious Diseases. *In:* OIE Quality Standard and Guidelines for Veterinary Laboratories: Infectious Diseases. OIE, Paris, France, 1–31.

15. WRIGHT P.F. (1998). International standards for test methods and reference sera for diagnostic tests for antibody detection. *Rev. sci. tech. Off. int. Epiz.*, **17,** 527–533.

16. WRIGHT P.F., NILSSON E., VAN ROOIJ E.M.A., LELENTA M. & JEGGO M.H. (1993). Standardization and validation of enzyme-linked immunosorbent assay techniques for the detection of antibody in infectious disease diagnosis. *Rev. sci. tech. Off. int. Epiz.*, **12**, 435–450.

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

*
* *