CHAPTER 1.1.6.

PRINCIPLES AND METHODS OF VALIDATION OF DIAGNOSTIC ASSAYS FOR INFECTIOUS DISEASES

INTRODUCTION

Validation is a process that determines the fitness of an assay[1]*, which has been properly developed, optimised and standardised, for an intended purpose. All diagnostic assays* (*laboratory and field assays*) *should be validated for the species in which they will be used. Validation includes estimates of the analytical and diagnostic performance characteristics of a test. In the context of this chapter, an assay that has completed the first three stages of the validation pathway* (*see Figure 1 below*), *including performance characterisation, can be designated as "validated for the original intended purpose*(*s*)*". To maintain a validated assay status, however, it is necessary to carefully monitor the assay's performance under conditions of routine use, often by tracking the behaviour of assay controls over time. This ensures that the assay, as originally validated, consistently maintains its performance characteristics. Should it no longer produce results consistent with the original validation data, the assay may be rendered unfit for its intended purpose*(*s*)*. Thus, a validated assay is continuously assessed to assure it maintains its fitness for purpose through both the assessment of results of the assay controls included with each run and through on-going assessment during routine use in the targeted population.*

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, contributing to eradication of 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 more intended purposes by optimising its performance characteristics for each purpose, e.g. setting diagnostic sensitivity (*DSe*) *high, with associated lower diagnostic specificity* (*DSp*) *for a screening assay, or conversely, setting DSp high with associated lower DSe for a confirmatory assay.*

The ever-changing repertoire of new and unique diagnostic reagents coupled with many novel assay platforms and protocols has precipitated discussions about how to properly validate these assays. It is no longer sufficient to offer simple examples from serological assays, such as the enzyme-linked immunosorbent assay, to guide assay developers in validating the more complex assays, such as nucleic acid detection tests. In order to bring coherence to the validation process for all types of assays, this chapter focuses on the criteria that must be fulfilled during assay development and validation of all assay types. The inclusion of assay development as part of the assay validation process may seem counterintuitive, but in reality, three of the required validation criteria (*definition of intended purpose*[*s*], *optimisation, and standardisation*) *that must be assessed in order to achieve a validated assay, comprise steps in the assay development process. Accordingly the assay development process seamlessly leads into an assay validation pathway, both of which require fulfilment of validation criteria. Further, more detailed guidance is provided in a series of Recommendations for validation of diagnostic tests*[2] *that are tailored for several fundamentally different types of assay* (*e.g. detection of nucleic acids, antibodies, or antigens*) *and provide more information on specific issues related to the validation of diagnostic assays. For*

1 "Assay," "test method," and "test" are synonymous terms for purposes of this chapter, and therefore are used interchangeably.

2 Available at: http://www.oie.int/fileadmin/Home/eng/Health_standards/tahm/3.6.00_INTRODUCTION.pdf

specific information for wildlife species, refer to Chapter 3.6.7 Principles and methods for the validation of diagnostic tests for infectious diseases applicable to wildlife*. The information provided in chapter 3.6.7, which is specific to wildlife species, might also be useful for domestic animal test validation, for example, where the number or availability of samples is limited.*

PRELIMINARY CONSIDERATIONS IN ASSAY DEVELOPMENT AND VALIDATION

All laboratories should comply with the requirements of Chapters 1.1.1 (*Aquatic Manual*) or 1.1.5 (*Terrestrial Manual*) on *Quality management in veterinary testing laboratories*. This will minimise the influence of factors that do not depend on the test itself such as instrumentation, operator error, reagent choice (both chemical and biological) and calibration, reaction vessels and platforms, water quality, pH and ionicity of buffers and diluents, incubation temperatures and durations, and errors in the technical performance of the assay.

The first step in assay development is to define the purpose of the assay, because this guides all subsequent steps in the validation process. Assay validation criteria are the characterising traits of an assay that represent decisive factors, measures or standards upon which a judgment or decision may be based. 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 categories such as: (a) the sample – individual or pooled, matrix composition, and host/organism interactions affecting the target analyte quantitatively or qualitatively; (b) the assay system – physical, chemical, biological and operator-related factors affecting the capacity of the assay to detect a specific analyte in the sample; and (c) the test result interpretation – the capacity of a test result, derived from the assay system, to predict accurately the status of the individual or population relative to the purpose for which the assay is applied.

Selection, collection, preparation, preservation and management of samples are critical variables in design and development of an assay to ensure valid test results. Other variables such as transport, chain of custody, tracking of samples, and the laboratory information management system are also key sources of variation/error that become especially important when the assay is implemented for routine testing. Integrity of experimental outcomes during assay development and validation is only as good as the quality of the samples used. Anticipating the factors that can negatively impact sample quality must precede launching an assay validation effort. Reference samples used in assay development and validation should be in the same matrix that is to be used in the assay (e.g. serum, tissue, whole blood) and representative of the species to be tested by the assay. The reference materials should appropriately represent the range of analyte concentration to be detected by the assay. Information on sample collection, preparation, preservation, management, and transport is available in chapters 1.1.2 and 1.1.3 of the *Terrestrial Manual*.

The matrix in which the targeted analyte is found (serum, faeces, tissue, etc.) may contain endogenous or exogenous inhibitors that prevent some assays from working. This is of particular concern for enzyme-dependent tests such as polymerase chain reaction (PCR) or enzyme-linked immunosorbent assay (ELISA). Other factors that affect the concentration and composition of the target analyte (particularly antibody) in the sample may be 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). Non-host factors, such as contamination or deterioration of the sample, also potentially affect the ability of the assay to detect the specific targeted analyte in the sample. It is also important that biological reagents are free of extraneous agents that might otherwise lead to erroneous results.

THE CRITERIA OF ASSAY DEVELOPMENT AND VALIDATION

Assay performance is affected by many factors beginning with optimisation of the assay. After initial optimisation for an intended purpose, characteristics of the performance of the assay will be tested. The assay may need additional optimisation or may be found to be fit for purpose based on the results of the validation work.

Criteria for Assay Development and Validation

i) Definition of the intended purpose(s)
ii) Optimisation
iii) Standardisation
iv) Repeatability
v) Analytical sensitivity
vi) Analytical specificity

vii) Thresholds (cut-offs)
viii) Diagnostic sensitivity
ix) Diagnostic specificity
x) Reproducibility
xi) Fitness for intended purpose(s)

A. ASSAY DEVELOPMENT PATHWAY

1. Definition of the intended purpose(s) for an assay

The OIE *Standard for Management and Technical Requirements for Laboratories Conducting Tests for Infectious Diseases* (World Organisation for Animal Health, 2008)[3] states that test methods and related procedures must be appropriate for specific diagnostic applications in order for the test results to be of relevance. In other words, the assay must be 'fit for purpose'. The qualitative and quantitative assessment of capacity of a positive or negative test result to predict accurately the infection or exposure status of the animal or population of animals is the ultimate consideration of assay validation. This capacity is dependent on development of a carefully optimised and standardised (Section A.2.5) assay that, through accrual of validation data, provides confidence in the assay's ability to perform according to the intended purpose. In order to ensure that test results provide useful diagnostic inferences about animals or populations with regard to the intended purpose, the validation process encompasses initial development and assay performance documentation, as well as on-going assessment of quality control and quality assurance 5mes. Figure 1 shows the assay validation process, from assay design through the development and validation pathways to implementation, deployment, and maintenance of the assay

The first step of assay development is selection of an assay type that is appropriate and that likely can be validated for a particular use (fitness for purpose).

The most common purposes are to:

1) Contribute to the demonstration of freedom from infection in a defined population (country/zone/compartment/herd) (prevalence apparently zero):

   1a) 'Free' with and/or without vaccination,

   1b) Re-establishment of freedom after outbreaks

2) Certify freedom from infection or presence of the agent in individual animals or products for trade/movement purposes.

3) Contribute to the eradication of disease or elimination of infection from defined populations.

4) Confirm 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).

These purposes are broadly inclusive of many narrower and more specific applications of assays (see the Recommendations for validation of diagnostic tests: footnote 2) for each assay type for details). Such specific applications and their unique purposes need to be clearly defined within the context of a fully validated assay.

Further to the intended purpose, the assay needs to be defined in terms of target animal species, target pathogen(s) or condition, and sampling matrix.

3 This is a specific interpretation of the more generally stated requirements of the ISO/IEC 17025:2005 international quality standard for testing laboratories (2005). This publication further states that for a test method to be considered appropriate, it must be properly validated and that this validation must respect the principles outlined in this document, the OIE Validation Standard.

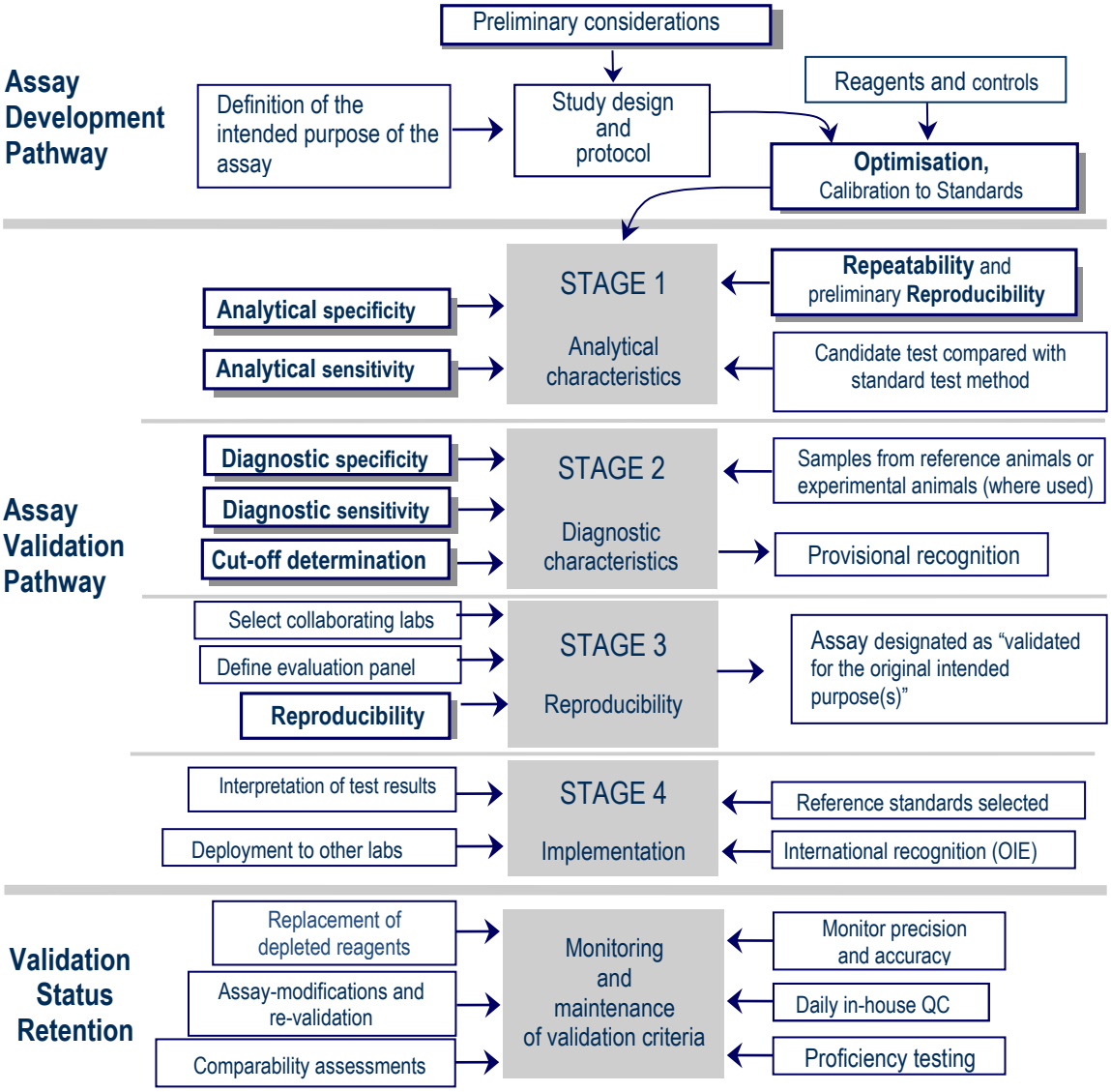

Figure 1. *The assay development and validation pathways with assay validation criteria highlighted in bold typescript within shadowed boxes.*

2. Assay development – the experimental studies

2.1. Test method design and proof of concept

Prior knowledge, thought and planning need to go into designing all steps of a new assay destined for validation, or an existing assay that is being modified. Assistance is offered in the Recommendations for validation of diagnostic tests (see footnote 2), which cover best practices for development and validation of assays for detection of various analytes (e.g. antibody, antigen, and nucleic acid detection).

Development of any assay is dependent on analyte reference samples that reflect the target analyte, the matrix in which the analyte is found, and the population for which the assay is intended to be used. The reference samples may be sera, fluids (including meat juices) or tissues that contain the analyte of interest or a genomic construct consistent with the target analyte. These reference materials are used in experiments conducted throughout the development process and carried over into the validation of the assay.

2.2. Operating range of the assay

The operating range of an assay is the interval of analyte concentrations or titres over which the method provides suitable accuracy and precision. Accuracy is the closeness of a test value to the expected (true) value (mean or median) for a reference standard reagent of known concentration or titre. Precision[4] is the degree of dispersion (variance, standard deviation or coefficient of variation) within a series of measurements of the same sample tested under specified conditions. During development of the assay, the lower and upper limits of the operating range are determined. To formally determine this range, a high positive reference sample is selected (ideally, this sample will be the same one from among the three samples described under "Optimisation" below). This high positive sample is serially diluted to extinction of the assay's response in an analyte-negative matrix of the same constitution as the sample matrix from animals in the population targeted by the assay. The results are plotted as a 'response-curve', with the response (e.g. optical density, cycle threshold, counts, etc.) a function of analyte concentration (amount). The curve establishes the working range of the assay. If the range is found to be unacceptable for the intended purpose, additional optimisation may be needed. The typical calibration curve for most assays is sigmoidal in shape. The data are transformed to approximate a linear relationship between response and concentration using a suitable algorithm (Findlay & Dillard, 2007).

2.3. Standardisation and optimisation

Optimisation is the process by which the most important physical, chemical and biological parameters of an assay are evaluated and adjusted to ensure that the performance characteristics of the assay are best suited to the intended application. It is useful to select at least three well-defined reference samples, representing the analyte ranging from high positive to negative (e.g. high, low positive and negative). These samples ideally should represent known infected and uninfected animals from the population that will become the target of the assay. Obtaining such reference samples, however, is not always possible, particularly for nucleic acid and antigen detection assays. The alternative of preparing reference samples spiked with cultured agents or positive sera is inferior as these samples do not truly represent the naturally occurring matrix-agent interaction (see also Chapter 3.6.6 *Selection and use of reference samples and panels*). When no other alternative exists, spiking a sample with a known amount of the analyte or agent derived from culture, or diluting a high positive serum in negative serum of the same species may be all that is available. In either case, it is imperative that the matrix, into which the analyte is placed or diluted, is identical to, or resembles as closely as possible the samples that ultimately will be tested in the assay. Ideally, reference samples have been well characterised by one or preferably at least two alternate methodologies. These samples can be used in experiments to determine if the assay is able to distinguish between varying quantities of analyte, distinguish the target from closely related analytes, and for optimising the reagent concentrations and perfecting the protocol. In principle, for all assay types, it is highly desirable to prepare and store a sufficient amount of each reference sample in aliquots for use in every run of the candidate assay as it is evaluated through the entire development and validation process. Switching reference samples during the validation process introduces an intractable variable that can severely undermine interpretation of experimental data and, therefore, the integrity of the development and validation process.

The labour-intensive process of optimising an assay is fundamental and critical to achieving a reliable and predictable assay performance. Scientific judgment and use of best scientific practices, as provided in the Recommendations for validation of diagnostic tests (see footnote 2), are recommended to guide optimisation of all elements of assay development and validation. The approach outlined provides a solid foundation for development of a reliable assay. Often, prototype assays are developed using reagents and equipment at hand in the laboratory. However, if the assay is intended for routine diagnostic use in multiple laboratories, standardisation becomes critical. Every chemical and buffer formulation must be fully described. All reagents must be defined with respect to purity and grade (including water). Acceptable working ranges must be established and documented for parameters such as pH, molarity, etc. Standards for quality, purity, concentration and reactivity of biologicals must be defined. Shelf lives and storage conditions must also be considered for both chemicals and biologicals. Acceptable ranges for reaction times and temperatures also need to be established. Essential equipment critical to assay performance must be described in detail, including operational specifications and calibration. Process (quality) control should be an integral part of optimisation and considered from the very beginning rather than, as is often the case at the end of assay development. In addition to the above, downstream aspects such as data capture, data manipulation and

4 Laboratory sources of variation that affect assay precision include: 1) within a single test run, 2) between concurrent runs, 3a) between assay runs at different times in the same day or on different days under similar conditions, 3b) between assay runs on different days with different operators, 4) between laboratories. In this chapter, categories 1–3 are estimates of repeatability, and category 4 is synonymous with reproducibility.

interpretation may also require standardisation and optimisation. Finally, all of these parameters, once optimised, must be fully described in the test method protocol.

During optimisation of an assay, it is important to take note of procedural steps and assay parameters that have a narrow range in which the assay performs optimally, as these are the critical points that ultimately affect an assay's reliability (see Section A.2.7). For some assay types, specific steps in the procedure may have more impact than other steps on the final assay performance (see Section B.5 below and Chapter 3.6.8 *Comparability of assays after minor changes in a validated test method* for additional information on establishing comparability when reagents or processes are changed).

A variety of analyte reference samples and other process controls that are routinely included in any assay system are identified in the following sections. These provide critical assay monitoring functions that require special attention during assay optimisation. In addition, attention must be paid to the proper preparation and storage of all biological reagents and reference materials to ensure stability (see Chapter 1.1.2).

2.4. Inhibitory factors in sample matrix

Each different matrix to be used in an assay must be used in the validation process. Some sample matrices include inhibitory factors that interfere with the performance of specific types of assays. Serum, particularly if haemolysed, may contain factors toxic to the cells used in viral neutralisation assays, while endogenous substances found in some tissues and fluids can interfere with or inhibit ligand-binding and enzymatic-based assays such as ELISAs. Faeces, autolysed tissues and semen samples tend to contain more interfering substances and are therefore more problematic for assay performance than are serum, blood or fresh tissues.

2.5. Robustness

Robustness refers to an assay's capacity to remain unaffected by minor variations in test situations that may occur over the course of testing in a single laboratory. Assessment of robustness should begin during assay development and optimisation stages. The deliberate variations in method parameters may be addressed in experiments after optimal conditions for an assay are established. However, when multi-factorial titrations of reagents are used for optimising the assay, indications of a compromised robustness may surface. If slight differences in conditions or reagent concentrations cause unacceptable variability, the assay most likely will not be robust. Early knowledge of this situation elicits a critical decision point for determining whether to continue with validation of the assay would be worthwhile, because if an assay is not robust within one laboratory under rather ideal conditions, it is unlikely to be reproducible when transferred to other laboratories.

The factors most likely to affect assay robustness include pH, temperature batch of reagents or brand of microtitre plates and aqueous or organic matrix factors (Dejaegher & Vander Heyden, 2006). Once optimisation is complete, the robustness of the assay becomes part of the assessment of repeatability

2.6. Calibration of the assay to standard reagents

2.6.1. International and national reference standards

Ideally, OIE or other international reference standards, containing a known concentration or titre of analyte, are the reagents to which all assays are standardised (see OIE Guide 3[5] and also chapter 3.6.6). Such standards are prepared and distributed by OIE Reference Laboratories or other international reference laboratories. National reference standards are calibrated by comparison with an international reference standard whenever possible; they are prepared and distributed by a national reference laboratory. In the absence of an international reference standard, a national reference standard becomes the standard of comparison for the candidate assay. These standards are highly characterised through extensive analysis, and preferably the methods for their characterisation, preparation, and storage have been published in peer-reviewed publications.

2.6.2. In-house standard

An in-house reference standard generally should be calibrated against an international or national standard. In the absence of either of these calibrators and to the extent possible, the in-house standard is highly characterised in the same manner as international and national

5 Available at: http://www.oie.int/fileadmin/Home/eng/Our_scientific_expertise/docs/pdf/GUIDELINE_3_REF_STANDARDS_ANG.pdf

standards (see chapter 3.6.6). This local in-house standard therefore becomes the best available standard, and is retained in sufficient aliquotted volumes for periodic use as the standard to which working standards are calibrated.

2.6.3. Working standard

One or more working standards, commonly known as analyte or process controls, are calibrated to an international, national, or in-house standard, and are prepared in large quantities, aliquotted and stored for routine use in each diagnostic run of the assay.

2.7. "Normalising" test results to a working standard

Due to the inherent variation in raw test results that are often observed between test runs of the same assay or among laboratories using the same or similar assays, it is almost impossible to compare directly (semi-) quantitative data. To improve markedly the comparability of test results both within and between laboratories, one or more working standard reagent(s) are included in each run of an assay. Raw test values for each test sample can then be converted to units of activity relative to the working standard(s) by a process called 'normalisation'. The 'normalised' values may be expressed in many ways, such as a per cent of a positive control (e.g. in an ELISA), or as the estimated concentration or titre of an analyte derived from a standard curve. It is good practice to include working standards in all runs of the assay during assay development and validation because this allows 'normalisation' of data, which provides a valid means for direct comparison of results between runs of an assay. It is mandatory to control the (absolute) variation of the normalisation standards as otherwise normalisation can introduce a bias. For more information, see Chapters 3.6.1 *Development and optimisation of antibody detection assays*, 3.6.2 *Development and optimisation of antigen detection assays* and 3.6.3 *Development and optimisation of nucleic acid detection assays*.

2.8. Preliminary study of the repeatability

Assessment of repeatability should begin during assay development and optimisation stages. Early knowledge of this situation elicits a critical decision point for determining whether it is worthwhile to continue with validation of the assay.

Repeatability is further verified during Stage 1 of assay validation (Section B.1.1). When the optimised test is run under routine laboratory or field conditions (Stage 4 of assay validation), repeatability is continually monitored as part of process control procedures for the duration of the life of the assay (see Section B.5.1).

B. ASSAY VALIDATION PATHWAY

"Validation" is a process that determines the fitness of an assay that has been properly developed, optimised and standardised for an intended purpose(s). Validation includes estimates of the analytical and diagnostic performance characteristics of a test. In the context of this document, an assay that has completed the first three stages of the validation pathway (Figure 1), including performance characterisation, can be designated as "validated for the original intended purpose(s)".

1. Stage 1 – Analytical performance characteristics

Ideally, the design of studies outlined in the following sections should be done with assistance of a statistician and a disease expert to ensure that the sample size and experimental approach are valid. It is possible to design experiments that efficiently provide information on likely within- and between-laboratory sources of variation in assay precision (see footnote 5 in Section A.2.2, above), which will define the performance characteristics of the assay. The choice of organisms, strains or serotypes to assess analytical sensitivity and specificity should reflect current knowledge and therefore inform the best possible experimental design for targeting specific analytes.

1.1. Repeatability

Repeatability is the level of agreement between results of replicates of a sample both within and between runs of the same test method in a given laboratory. Repeatability is estimated by evaluating variation in results of replicates. The number of replicates should preferably be determined in consultation with a statistician with a suggested minimum of three samples representing analyte activity within the operating range of the assay. Each of these samples is then aliquoted into the appropriate number of individual vessels as identical replicates of the original sample containing the original analyte and matrix concentration (see chapter 3.6.6). Each replicate is then run through all steps of the

assay, including creating the working dilution, as though it were a test sample derived from the population targeted by the assay. It is not acceptable to prepare a final working dilution of a sample in a single tube from which diluted aliquots are pipetted into reaction vessels, or to create replicates from one extraction of nucleic acid rather than to extract each replicate before dilution into the reaction vessels. Such 'samples' do not constitute valid replicates for repeatability studies. Between-run variation is determined by using the same samples in multiple runs involving two or more operators, done on multiple days. The variation in replicate results can be expressed as standard deviations, coefficients of variation (standard deviation ÷ mean of replicates), or other possible options (see chapter 3.6.4 *Measurement uncertainty* for assessments of repeatability).

1.2. Analytical specificity

Analytical specificity (ASp) is the ability of the assay to distinguish the target analyte (e.g. antibody, organism or genomic sequence) from non-target analytes, including matrix components. The assessment is qualitative and the choice and sources of sample types, organisms and sequences for the ASp evaluation should reflect test purpose and assay type. See chapters 3.6.1, 3.6.2 and 3.6.3 for guidance for antibody, antigen and nucleic acid assays, respectively. ASp is documented during Stage 1 validation, and cross-reactions identified. Cross-reactivity (ASp less 100%) may be acceptable depending on the proposed use of the assay. The impact of cross-reactivity is further documented during Stage 2 (establishment of DSp) and assessed at Stage 4 implementation.

1.2.1. Selectivity

Selectivity refers to the extent to which a method can accurately quantify the targeted analyte in the presence of: 1) interferents such as matrix components (e.g. inhibitors of enzymes in the reaction mix); 2) degradants (e.g. toxic factors); 3) nonspecific binding of reactants to a solid phase (e.g. conjugate of an ELISA adsorbed to well of microtiter plate); 4) antibodies to vaccination that may be confused with antibodies to active infection. Such interferents may cause falsely reduced or elevated responses in the assay that negatively affect its analytical specificity. Vessman *et al.* (2001) is a useful overview of selectivity as defined for analytical chemistry from which a modification described herein was deduced for application to diagnostic tests.

1.2.2. Exclusivity

Exclusivity is the capacity of the assay to detect an analyte or genomic sequence that is unique to a targeted organism, and excludes all other other known organisms that are potentially cross-reactive. This would also define a confirmatory assay.

1.2.3. Inclusivity

Inclusivity is the capacity of an assay to detect several strains or serovars of a species, several species of a genus, or a similar grouping of closely related organisms or antibodies thereto. It characterises the scope of action for a screening assay.

1.3. Analytical sensitivity

The limit of detection (LOD) is a measure of the analytical sensitivity (ASe) of an assay. The LOD is the estimated amount of analyte in a specified matrix that would produce a positive result at least a specified percent of the time. Typically, estimated LOD will be based on spiking of the analyte into the target matrix. The choice of analyte(s) (e.g. species, strains) is part of the ASe definition and should be reported properly. These experiments may be designed for precise and accurate estimation of the probability point (e.g. 50% or 100%), but in some circumstances a conservative estimate of the LOD (e.g. 100%) may be acceptable. For example, in a titration using tenfold dilutions all replicates at all dilutions might show either 100% or 0% response. There are two choices at that point. The last dilution showing 100% response may be accepted as a conservative estimate of the lower limit of detection. A more accurate estimate may be obtained by a second stage experiment using narrower intervals in the dilution scheme focusing on the region between 100% and 0%. Methods for statistical evaluation of LOD data are in the Chapter 3.6.5 *Statistical approaches to validation*.

1.4. Analytical accuracy of adjunct tests or procedures

Some test methods or procedures may be qualified for use as analytical tools in the diagnostic laboratory. These usually are secondary adjunct tests or procedures that are applied to an analyte that has been detected in a primary assay. The purpose of such analytical tools is to further characterise

the analyte detected in the primary assay. Examples of such adjunct tests include virus neutralisation to type an isolated virus, and molecular sequencing.

Such adjunct tests must be validated for analytical performance characteristics (Sections A.2 through B.1.3, above). However, they differ from diagnostic tests because they do not require validation for diagnostic performance characteristics (Sections B.2 through B.4, below) if their results are not used to establish a final diagnosis with regard to the intended purpose. The analytical accuracy of these tools may be defined by comparison with a reference reagent standard, or by characteristics inherent in the tool itself (such as endpoint titration). In all of these examples, the targeted analyte is further characterised quantitatively or qualitatively by the analytical tool.

2. Stage 2 – Diagnostic performance of the assay

Estimates of DSe (proportion of samples from known infected reference animals that test positive in an assay) and DSp (the proportion of samples from known uninfected reference animals that test negative in an assay) are the primary performance indicators established during validation of an assay. These estimates are the basis for calculation of other parameters from which inferences are made about test results (e.g. predictive values of positive and negative test results). Therefore, it is imperative that estimates of DSe and DSp are as accurate as possible. Ideally, they are derived from testing a panel 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. An estimate of the area under the receiver operating characteristic (ROC) curve is a useful adjunct to DSe and DSp estimates for a quantitative diagnostic test because it assesses its global accuracy across all possible assay values (Greiner *et al.,* 2000; Zweig & Campbell, 1993). This approach is described in chapter 3.6.5.

The designated number of known positive and known negative samples will depend on the likely values of DSe and DSp of the candidate assay and the desired confidence level for the estimates (Table 1 and Jacobson, 1998). Table 1 provides two panels of the theoretical number of samples required, when either a 5% or 2% error is allowed in the estimates of DSe or DSp. Many samples are required to achieve a high confidence (typically 95%) in the estimates of DSe and DSp when a small error margin in the estimate is desired. For example comparison of

Table 1. *Theoretical number of samples from animals of known infection status required for establishing diagnostic sensitivity (DSe) and specificity (DSp) estimates depending on likely value of DSe or DSp and desired error margin and confidence*

Estimated DSe or DSp	**2%** error allowed in estimate of DSe and DSp			**5%** error allowed in estimate of DSe and DSp		
	Confidence			Confidence		
	90%	95%	99%	90%	95%	99%
90%	610	864	1493	98	138	239
92%	466	707	1221	75	113	195
94%	382	542	935	61	87	150
95%	372	456	788	60	73	126
96%	260	369	637	42	59	102
97%	197	279	483	32	45	77
98%	133	188	325	21	30	52
99%	67	95	164	11	15	26

a 2% vs 5% error for a likely DSe or Dse of 90% and 95% confidence shows a considerable increase (864 *vs* 138) in the number of samples required. Logistical and financial limitations may require that less than the statistically required sample size will be evaluated, in which case the confidence interval calculated for DSe and DSp will indicate less diagnostic confidence in the results. Sample size also may be limited by the fact that reference populations and OIE reference standards may be lacking (see chapter 3.6.5 for further details). It may, therefore, be necessary to use a sub-optimal number of samples initially. It is, however, highly desirable to enhance confidence and reduce error margin in the DSe and DSp estimates by adding more samples (of equivalent status to the original panel) as they become available.

The following are examples of reference populations and methodologies that may aid in determining performance characteristics of the test being validated.

2.1. Reference animal populations

Ideally, selection of reference animals requires that important host variables in the target population are represented in animals chosen for being infected with or exposed to the target agent, or that have never been infected or exposed. The variables to be noted include but are not limited to species, age, sex, breed, stage of infection, vaccination history, and relevant herd disease history (for further details see chapter 3.6.6).

2.1.1. Negative reference samples

True negative samples, from animals that have had no possible infection or exposure to the agent, may be difficult to locate. It is often possible to obtain these samples from countries or zones that have eradicated or have never had the disease in question. Such samples may be useful as long as the targeted population for the assay is sufficiently similar to the sample-source population.

2.1.2. Positive reference samples

It is generally problematic to find sufficient numbers of true positive reference animals, as determined by isolation of the pathogen. It may be necessary to resort to samples from animals that have been identified by another test of sufficiently high accuracy, such as a validated nucleic acid detection assay. The candidate test is applied to these reference samples and results (positive and negative) are cross-classified in a 2 × 2 table. This has been called the "gold standard model" as it assumes the reference standard is perfect. A sample calculation is shown in Table 2 in Section B.2.5).

2.2. Samples from animals of unknown status

When the so-called reference standard is imperfect, which is the rule with any diagnostic tests, estimates of DSe and DSp for the candidate assay based on this standard will be flawed. A way to overcome this problem is to perform a latent class analysis of the joint results of the two tests assuming neither test is perfect.

Latent-class models do not rely on the assumption of a perfect reference test but rather estimate the accuracy of the candidate test and the reference standard with the joint test results (Branscum *et al.,* 2005; Enøe *et al.,* 2000; Georgiadis *et al.,* 2003; Hui & Walter, 1980). If a Bayesian latent class analysis is used, prior knowledge about the performance of the reference test and the candidate test can be incorporated into the analysis.

Because these statistical models are complex and require critical assumptions, statistical assistance should be sought to help guide the analysis 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 (see chapter 3.6.5 for details).

2.3. Experimentally infected or vaccinated reference animals

Samples obtained sequentially from experimentally infected or vaccinated animals are useful for determining the kinetics of antibody responses or the presence/absence of antigen or organisms in samples from such animals. However, multiple serially acquired pre- and post-exposure results from individual animals are not acceptable for establishing estimates of DSe and DSp because the statistical requirement of independent observations is violated. Single time-point sampling of individual experimental animals can be acceptable (e.g. one sample randomly chosen from each animal). Nevertheless it should be noted that for indirect methods of analyte detection, 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 (Jacobson, 1998). 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. In cases when the near-impossibility of obtaining suitable reference samples from naturally exposed animals necessitates the use of samples from experimental animals for validation studies, the resulting DSe and DSp measures should be considered as less than ideal estimates of the true DSp and DSe.

2.4. Cut-off (threshold) determination

To obtain DSe and DSp estimates of the candidate assay, which is measured on a continuous scale, the test results first must be reduced to two (positive or negative) or three (positive, intermediate [doubtful] or negative) categories of test results. This is accomplished by insertion of one or two cut-off points (threshold or decision limits) on the scale of test results. The selection of the cut-off(s) should reflect the intended purpose of the assay and its application, and must support the required DSe and DSp of the assay. Options and descriptive methods for determining the best way to express DSe and DSp are available (Branscum *et al.*, 2005; Georgiadis *et al.,* 2003; Greiner *et al.,* 1995; Greiner *et al.,* 2000; Jacobson, 1998; Zweig & Campbell, 1993; and chapter 3.6.5). If considerable overlap occurs in the distributions of test values from known infected and uninfected animals, it is impossible to select a single 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) (Greiner *et al.,* 1995).

The main difficulty in establishing cut-offs based on diagnostic performance characteristics is the lack of availability of the required number of well-characterised samples. Alternatives are discussed in Section B.2.6 on provisional acceptance of an assay during accrual of data to enhance estimates of DSe and DSp.

2.5. Calculation of DSe and DSp based on test results of reference samples

A typical method for determining DSe and DSp estimates is to test the reference samples in the new assay, and cross tabulate the categorical test results in a 2 × 2 table. In a hypothetical example, assume the test developer has selected a sample size for DSe and DSp for the new assay under the assumption that the most likely values are 97% (DSe) and 99% (DSp), respectively, with a desired confidence of 95% for both estimates. The desired error margin in the estimates was set at 2%. Table 1 indicates that 279 samples from known infected animals are required for the DSe assessment, and 95 known negative samples are needed for establishing the DSp estimate. The samples were then run in the new assay. Table 2 is a hypothetical set of results from which DSe and DSp estimates have been obtained.

Table 2. *Diagnostic sensitivity and specificity estimates calculated from hypothetical set of results for samples tested from known infected and non-infected populations*

		Number of reference samples required*	
		Known positive (279)	Known negative (95)
Test results	Positive	270 TP	FP 7
	Negative	FN 9	TN 88
		Diagnostic sensitivity* TP/(TP + FN) **96.8%** (94.0 – 98.5%)**	Diagnostic specificity* TN/(TN + FP) **92.6%** (85.4 – 97.0%)**

*Based on Table 1 for an assay with the following parameters:
1) Prior to testing, estimated DSe of 97% and DSp of 99%
2) 95% = required confidence in DSe and DSp estimates
3) 2% = Error margin in the estimates of DSe and DSp
TP and FP = True Positive & False Positive, respectively
TN and FN = True Negative and False Negative, respectively
**95% exact binomial confidence limits for DSe and DSp calculated values
(see chapter 3.6.5 for information on confidence limits)

In this example, the DSe estimate is as anticipated, but the DSp is much lower (92%) than the anticipated value of 99%. As a consequence, the width of the confidence interval for DSp is greater than expected. Re-inspection of Table 1 indicates that 707 samples are necessary to achieve an error margin of ± 2% at a DSP of 92% but such an increase in sample size might not be feasible (see chapter 3.6.5 for further details).

2.6. Provisional assay recognition[6]

There are situations where it is not possible or desirable to fulfil Stage 2 of the Validation Pathway because appropriate samples from the target population are scarce and animals are difficult to access (such as for transboundary infectious diseases or wildlife diseases).

Experience has shown that the greatest obstacle for continuing through Stage 2 of the Validation Pathway is the number of defined samples required to calculate DSe and DSp. The formula is well known and tables are available for determining the number of samples required to estimate various levels of DSe and DSp, depending on the desired error margin and the level of confidence in the estimates (Table 1 and Jacobson, 1998). The formula assumes that the myriad of host/organism factors that may affect the test outcome are all accounted for. Since that assumption may be questionable, the estimated sample sizes are at best minimal. For a disease that is not endemic or widespread, it may be impossible, initially, to obtain the number of samples required, but over time, accrual of additional data will allow adjustment of the cut-off (threshold) or if no adjustment is needed, enhance confidence in the estimates.

Provisional recognition defines an assay that has been assessed through Stage 1 for critical assay benchmark parameters (ASe, ASp and repeatability) with, in addition, a preliminary estimate of DSp and DSe based on a small select panel of well-characterised samples containing the targeted analyte and a preliminary estimate of reproducibility. This represents partial completion of Stage 2. Preliminary reproducibility estimates of the candidate assay could be done using the select panel of well-characterised samples to enhance provisional acceptance status for the assay. The candidate test method is then duplicated in laboratories in at least two different institutes, and the panel of samples is evaluated using the candidate assay in each of these laboratories, using the same protocol, same reagents as specified in the protocol, and comparable equipment. This is a scaled-down version of the reproducibility study in Stage 3 of assay validation. In following this procedure of provisional recognition the test protocol must not be varied.

Provisional recognition of an assay by state or national authorities means that the assay has not been evaluated for diagnostic performance characteristics. As such, the laboratory should develop and follow a protocol for adding and evaluating samples, as they become available, to fulfil this requirement. Ideally, this process should be limited to a specific timeframe in which such an accrual would be directed toward fulfilling Stages 2 and 3 of the validation pathway, and to particular situations (emergencies, minor species, no other test available, etc.)

3. Stage 3 – Reproducibility and augmented repeatability estimates

3.1. Reproducibility

Reproducibility is the ability of a test method to provide consistent results, as determined by estimates of precision, when applied to aliquots of the same samples tested in different laboratories, preferably located in distinct or different regions or countries using the identical assay (protocol, reagents and controls). To assess the reproducibility of an assay, each of at least three laboratories should test the same panel of samples (blinded) containing a suggested minimum of 20 samples, with identical aliquots going to each laboratory (see chapter 3.6.6). This exercise also generates preliminary data on non-random effects attributable to deployment of the assay to other laboratories. In addition, within-laboratory repeatability estimates are augmented by the replicates used in the reproducibility studies. Measurements of precision can be estimated for both the reproducibility and repeatability data (see chapter 3.6.4 for further explanation of the topic and its application).

For field tests, reproducibility should be evaluated under the conditions of intended use.

3.2. Designation of a validated assay

On completion of Stage 3 validation, assuming the earlier stages have been fully and satisfactorily completed, the assay may be designated as “validated for the original intended purpose”. Retention of this designation is dependent on continual monitoring of the assay performance, as described in Section 5.1.

6 Provisional recognition does not imply acceptance by the OIE. It does, however, recognise an informed decision of authorities at local, state, national or international levels of their conditional approval of a partially validated assay.

4. Stage 4 – Programme implementation

The successful deployment of an assay provides additional and valuable evidence for its performance according to the expectations. Moreover, the (true) prevalence of the diagnostic trait in the target population is an important factor that needs to be accounted for as described below.

4.1. Fitness for use

While this chapter deals with validation and fitness for purpose from a scientific perspective, it should also be noted that other practical factors might impact the utility of an assay with respect to its intended application. 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. For some diseases, multiple assays might be available for use in combination in disease control and surveillance programmes and hence, an assay's utility might need to be assessed by evaluating incremental changes in DSe, DSp and predictive values of the combined tests.

An inability to meet operational requirements of an assay also may make it unfit for its intended use. Such requirements may include performance costs, equipment availability, level of technical sophistication and interpretation skills, kit/reagent availability, shelf life, transport requirements, safety, biosecurity, sample throughput, turn-around times for test results, aspects of quality control and quality assurance, and whether the assay can practically be deployed to other laboratories. Test kits used in the field are highly desirable from an ease-of-use viewpoint, but because they are performed outside the confines of a controlled laboratory environment, they require added precautions to maintain fitness for purpose (Crowther *et al.,* 2006).

4.2. Interpretation of test results

Predictive values of test results: The positive predictive value (PPV) is the probability that an animal that has tested positive is in fact positive with regard to the true diagnostic status. The negative predictive value (NPV) is the probability that an animal that has tested negative is in fact negative with regard to the true diagnostic status.

Predictive values of test results are an application of Bayes' theorem and are calculated as follows:

$$PPV = \frac{P \times DSe}{P \times DSe + (1 - P) \times (1\text{-}DSp)}$$

and

$$NPV = \frac{(1 - P) \times DSp}{P \times (1 - DSe) + (1 - P) \times DSp}$$

Where:

PPV = Predictive value of a positive test result

NPV = Predictive value of a negative test result

P = Prevalence of infection

DSe = Diagnostic sensitivity

DSp = Diagnostic specificity

In contrast to DSe and DSp, predictive values are influenced by the true prevalence of the true diagnostic status of the target population. In other words, predictive values are not inherent characteristics of a specific diagnostic test, but are a function of its DSe and DSp and the local prevalence of infection in a defined population at a given point in time.

Predictive values are of great importance to field veterinarians for the interpretation of results. For example, a PPV of 0.9 means that an animal reacting positive to the test has 90% chance of being indeed infected and 10% probability of being a false positive.

The predictive value of a positive result also has great importance for the veterinary services in charge of the management of control or eradication programmes. If we consider the inverse of the PPV (i.e. 1/PPV) it gives the information on how much money is spent in the culling of true and false positives for each true positive animal detected by the surveillance activity. In other words, if the PPV V is 0.67, it means that two positive animals out of three are true positives and the remaining is a false positive. Since during the application of a control programme, the prevalence of infection is continually changing, the monitoring of the PPV is a way of evaluating the costs of the programme.

Furthermore, during the application of a control programme it is usually advisable to change the sensitivity of the tests employed, based on the variation of prevalence of infection in the target population and on the objective of the programme, the PPV may be used to make the changes in DSe and DSp based on economic considerations. In other words, when the need for a change in DSe and DSp of the test arises, a number of putative cut-offs may be set along the ROC curve of the test validation and the relevant values of DSe and DSp for each cut-off may be used to evaluate the expected cost for the culling of each infected animal.

If the purpose is establishing evidence for freedom from disease, the NPV is the more important measure. The NPV critically depends on DSe.

4.3. International recognition

Traditionally, assays have been recognised internationally by the OIE when they are designated as prescribed or alternate tests for trade purposes. This has often been based on evidence of their usefulness on a national, regional or international basis. For commercial diagnostic kits that have gone through the OIE procedure for validation and certification of diagnostic assays, the final step is listing of the test in the OIE Register. Tests listed in the Register are certified as fit for a specific purpose if they have completed Validation Stages 1, 2 and 3. The Register is intended to provide potential kit users with an informed and unbiased source of information about the kit and its performance characteristics for an intended purpose. The Register is available on the OIE website at: http://www.oie.int/en/our-scientific-expertise/certification-of-diagnostic-tests/the-register-of-diagnostic-tests/.

4.4. Deployment of the assay

Ultimate evidence of the usefulness of an assay is its successful application(s) in other laboratories and inclusion in national, regional and/or international control or surveillance programmes. Reference laboratories play a critical role in this process. In the natural progression of diagnostic and/or technological improvements, new assays will become the new standard method to which other assays will be compared. 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.

An assessment of the reproducibility should be repeated when the test is transferred from the development laboratory to the field, whether for use in local laboratories or in field applications. Predictable changes, e.g. extremes of temperature and levels of operator experience, should be assessed as additional sources of variation in assay results that may affect estimates of reproducibility.

5. Monitoring assay performance after initial validation

5.1. Monitoring the assay

To retain the status of a validated assay it is necessary to assure that the assay as originally validated consistently maintains the performance characteristics as defined during validation of the assay. This can be determined in a quality assurance programme characterised by carefully monitoring the assay's daily performance, primarily through precision and accuracy estimates for internal controls, as well as outlier tendencies. The performance can be monitored graphically by plotting measurements from assay controls in control charts[7]. Deviations from the expected performance should be investigated so corrective action can be taken if necessary. Such monitoring provides critical evidence that the assay retains its "validated" designation during the implementation phase of the assay. Reproducibility is assessed through external quality control programmes such as proficiency testing. Should the assay cease to produce results consistent with the original validation data, the assay would be rendered unfit

7 *Control chart:* A graphical representation of data from the repetitive measurement of a control sample(s) tested in different runs of the assay over time.

for its intended purpose. Thus, a validated assay must be continuously assessed to assure it maintains its fitness for purpose.

5.2. Modifications and enhancements – considerations for changes in the assay

Over time, modifications of the assay likely will be necessary to address changes in the intended purpose, analytes targeted (i.e. modification of the assay to adjust diagnostic performance) or technical modifications to improve assay efficiency or cost-effectiveness. For a change in intended purpose of the assay, then a revised validation from Stage 2 onwards is obligatory.

If the assay is to be applied in another geographical region and/or population, revalidation of the assay under the new conditions is recommended. Lineages or sub-lineages of an infectious agent, derived from animals in different geographic locations, are known to vary requiring revalidation of the assay for the specified target population. This is especially true for nucleic acid detection (NAD) systems as it is very common for point mutations to occur in many infectious agents (especially RNA viruses). Mutations, which may occur within the primer or probe sites can affect the efficiency of the assay and even invalidate the established performance characteristics. It is also advisable to regularly confirm the target sequence at the selected genomic regions for national or regional isolates of the infectious agents. This is especially true for the primer and probe sites, to ensure that they remain stable and the DSe and DSp for the assay are not compromised. Similar issues can arise with immunologically based assays for antigen or antibody.

A similar situation may occur with emergence of new subtypes of existing pathogens. In these circumstances, existing assays may need to be modified.

5.2.1. Technical modifications and comparability assessments

Technical modifications to a validated assay such as changes in instrumentation, extraction protocols, and conversion of an assay to a semi-automated or fully automated system using robotics will typically not necessitate full revalidation of the assay. Rather, a methods comparison study is done to determine if the relatively minor modification to the assay affected the previously documented performance characteristics of the assay. Comparability can be established by running the modified procedure and original procedure side-by-side, with the same panel of samples in both, over several runs. The panel chosen for this comparison should represent the entire operating range of both assays. If the results from the modified procedure and originally validated method are determined to be comparable in an experiment based on a pre-specified criterion, the modified assay remains valid for its intended purpose. See chapter 3.6.8 for description of experiments that are appropriate for comparability testing, and chapter 3.6.6 on reference sample panels.

5.2.2. Biological modifications and comparability assessments

There may be situations where changes to some of the biologicals used in the assay may be necessary and/or warranted. This may include changes to the test specimen itself (e.g. a change in tissue to be tested or perhaps testing of a different species altogether). It may include changes to reagents (e.g. the substitution of a recombinant antigen for a cell culture derived antigen or one antibody conjugate for another of similar immunological specificity in an ELISA). The difficulty in making any modification lies in determining whether the change requires a complete revalidation of the assay at both bench and field levels. At the very least, any modification requires that the appropriate Stage 1 'analytical requisites' be assessed. The more difficult decision relates to Stage 2 'diagnostic performance'. To assist here, the original (reference) assay should initially be compared to the modified (candidate) assay in a controlled trial using a defined panel of positive and negative diagnostic samples. See chapter 3.6.8 for a description of comparability assessment. If the comparability assessment does not suggest a change in diagnostic performance, the modified assay may be phased into routine use. If, on the other hand, differences in DSp and DSe are observed, the modified assay would require additional Stage 2 or field validation before being adopted.

5.2.3. Replacement of depleted reagents

When a reagent such as a control sample or working standard is nearing depletion, it is essential to prepare and repeatedly test a replacement before such a control is depleted. The prospective control sample should be included in multiple runs of the assay in parallel with the original control to establish their proportional relationship. It is important to change only one reagent at a time to avoid the compound problem of evaluating more than one variable.

5.3. Enhancing confidence in validation criteria

Because many host variables have an impact on the diagnostic performance of assays, it is highly desirable over time to increase the number of reference samples or samples suitable for latent class analysis. The sampling design, collection, transportation, and testing environment for the new samples should be the same as used for the original validation study. Increases in sample numbers improves the precision of the overall estimates of DSe and DSp, and may allow calculations of DSe estimates by factors such as age, stage of disease, and load of organisms. New data should be included annually in relevant test dossiers.

5.4. Verification of existing assays (in-house validation)

If a laboratory is considering the use of a validated commercial kit or a candidate assay based on published literature with validation data, some form of verification will be required to determine whether the assay complies with either the kit manufacturer's or the author's assertions, with respect to Stage 1 validation criteria, in the context of the intended application. This may require a limited verification of both ASp and ASe using available reference materials, whether they be external and/or locally acquired from the target population. Once the laboratory is confident that the assay is performing as described from an analytical perspective, then proceeding to a limited Stage 2 validation should be considered in the context of the intended application and target population before the assay is put into routine diagnostic use.

REFERENCES

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

CROWTHER J.R., UNGER H. & VILJOEN G.J. (2006). Aspects of kit validation for tests used for the diagnosis and surveillance of livestock diseases: producer and end-user responsibilities. *Rev. sci. tech. Off. int. Epiz*., **25** (3), 913–935.

DEJAEGHER B. & VANDER HEYDEN Y. (2006). Robustness tests. *LCGC Europe*, **19** (7), online at http://www.lcgceurope.com/lcgceurope/content/printContentPopup.jsp?id=357956

ENØE 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.

FINDLAY J.W.A. & DILLARD R.F. (2007). Appropriate calibration curve fitting in ligand binding assays. *AAPS J*., **9** (2): E260-E267. (Also on-line as *AAPS Journal* [2007]; **9** [2], Article 29 [http://www.aapsj.org]).

GEORGIADIS M., JOHNSON, W., GARDNER I. & SINGH R. (2003). Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Appl. Statist*., **52** (Part 1), 63–76.

GREINER M., SOHR D. & GÖBEL P. (1995). A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests. *J. Immunol. Methods*, **185**,123–132.

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.

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

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

VESSMAN J., STEFAN R., VAN STADEN J., DANZER K., LINDNER W., BURNS D., FAJGELJ A. & MULLER H. (2001). Selectivity in analytical chemistry. *Pure Appl. Chem.,* **73** (8), 1381–1386.

WORLD ORGANISATION FOR ANIMAL HEALTH (OIE) (2008). 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.

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