CHAPTER 1.1.4.

PRINCIPLES AND METHODS OF VALIDATION OF DIAGNOSTIC ASSAYS FOR INFECTIOUS DISEASES

INTRODUCTION

Validation is a process that determines the fitness of an assay, which has been properly developed, optimised and standardised, for an intended purpose. 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)"[1]. To maintain a validated assay status, however, it is necessary to carefully monitor the assay's daily performance, often by tracking the behaviour of internal 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. Thus, a validated assay is continuously assessed to assure it maintains its fitness for purpose through assessment of results of internal controls in each run of the assay.

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

The terms "**valid**" (adjective) or "**validity**" (noun) refer to whether estimates of test performance characteristics are unbiased with respect to the true parameter values. These terms are applicable regardless of whether the measurement is quantitative or qualitative.

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, 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 validation criteria that must be assessed in order to achieve a validated assay, comprise steps in the assay development process. Accordingly the assay development process seamlessly segues into an assay validation pathway, both of which contain validation criteria that must be fulfilled. This chapter also provides guidance for evaluation of each criterion through provision of best scientific practices contained in the chapter's appendices. The best practices are tailored for each of several fundamentally different types of assays (e.g. detection of nucleic acids, antibodies, or antigens).

1 **Validation** does not necessarily imply that test performance meets any minimum value or that the test has comparable performance to any comparative test, unless this has been specifically considered in the design of the test evaluation study.

DIRECT AND INDIRECT METHODS THAT REQUIRE VALIDATION

The diagnosis of infectious diseases is performed by direct and/or indirect detection of infectious agents. By direct methods, the particles of the agents and/or their components, such as nucleic acids, structural or non-structural proteins, enzymes, etc., are detected. The indirect methods demonstrate antibodies or cell-mediated immune responses induced by exposure to infectious agents or their components. The most common indirect methods of infectious agent detection are antibody assays such as classical virus neutralisation, antibody enzyme-linked immunosorbent assay (ELISA), haemagglutination inhibition, complement fixation, and the recently appearing novel methods, such as biosensors, bioluminometry, fluorescence polarisation, and chemoluminescence,

The most common direct detection methods are isolation or *in-vitro* cultivation of viable organisms, electron microscopy, immunofluorescence, immunohistochemistry, antigen-ELISA, Western immunoblotting, and nucleic acid detection systems (NAD). The NAD systems include nucleic-acid hybridisation (NAH), macro- and microarrays and the various techniques of nucleic acid amplification, such as the polymerase chain reaction (PCR), or the isothermal amplification methods, such as nucleic acid sequence-based amplification (NASBA), and invader or loop-mediated isothermal amplification (LAMP). NAD assays are rapidly becoming commonplace and in many cases replacing virus isolation and bacteria cultivation, particularly for the detection of agents that are difficult or impossible to culture. NAD tools are also used as a secondary means for highly specific identification of strains, groups, or lineages of organisms following isolation or culture of viruses, bacteria and parasites. Molecular diagnostics, such as PCR, do not require: a) the presence of replicating organisms, b) expensive viral isolation infrastructure, c) up to several weeks to achieve a diagnosis, or d) special expertise, which is often unavailable in many laboratories – all practical advantages. These methods have become relatively inexpensive, safe and user-friendly (1–4, 6, 7, 16, 20, 21). Various real-time PCR methods, nucleic acid extraction robots, and automated workstations for NAD, antibody, antigen, and agent detection have resulted in a large repetoire of high throughput, robust, very rapid and reliable assays. Although NAD systems often have the advantage of a greater diagnostic sensitivity and analytical specificity than infectious agent recovery or antigen-capture ELISA procedures, that advantage usually carries with it a greater challenge for validation of such assays.

PRELIMINARY CONSIDERATIONS IN ASSAY DEVELOPMENT AND VALIDATION

The first consideration is to define the purpose of the assay, because this guides all subsequent steps in the validation process. 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 – 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 analyte in question.

Assay validation criterion: a characterising trait of an assay; a decisive factor, measure or standard upon which a judgment or decision may be based.

The matrix in which the targeted analyte may reside (serum, faeces, tissue, etc.) may contain endogenous or exogenous inhibitors that prevent enzyme-dependent tests such as PCRs or ELISAs from working. Other factors that affect the concentration and composition of analyte (mainly 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 affect the ability of the assay to detect the specific targeted analyte in the sample.

Factors that interfere with the analytical performance of the assay system include instrumentation, operator error, reagent choice (both chemical and biological) and calibration, accuracy and acceptance limits of assay controls, reaction vessels and platforms, water quality, pH and ionicity of buffers and diluents, incubation temperatures and durations, and error introduced by detection of closely related analytes. It is also important that biological reagents are free of extraneous agents.

Factors that may negatively impact diagnostic performance of the assay are primarily associated with choice of reference sample panels from known infected/exposed or known uninfected animals selected for evaluating the diagnostic sensitivity (DSe) and diagnostic specificity (DSp) of the assay. This is particularly difficult because 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 confidence of test-result interpretation. For example, experienced serologists are aware that an assay, validated by using serum samples from northern European cattle, may not give valid results for the distinctive populations of cattle in Africa. Diagnostic performance of the assay is further complicated when sample panels of known infection status are not available, often because they

are impossible to obtain. In this situation, DSe and DSp may be estimated, in certain circumstances by use of latent class models (10, 14 and Appendix 1.1.4.5).

THE CRITERIA OF ASSAY DEVELOPMENT AND VALIDATION

Assay performance is affected by many factors that span from the earliest stages of assay development through the final stage of performance assessment when the test is applied to targeted populations of animals. An assay, therefore, cannot be considered validated unless a specific set of essential validation criteria (see accompanying box) have been tested and affirmed or fulfilled, either quantitatively or qualitatively (for detail on terms, see glossary in reference 25). Lack of attention to any one of these criteria will likely reduce the level of confidence that an assay is fulfilling the purpose(s) for which it is intended. The first four of these criteria typically are addressed during development of the assay (the Development Pathway), and the remaining eight are evaluated during the first three stages of assay validation (the Validation Pathway) as described below.

Assay validation criteria

1. Fitness for intended purpose(s)
2. Optimisation
3. Standardisation
4. Robustness
5. Repeatability
6. Analytical sensitivity
7. Analytical specificity
8. Thresholds (cut-offs)
9. Diagnostic sensitivity
10. Diagnostic specificity
11. Reproducibility
12. Ruggedness

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* (25) 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[2]. The 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 (Section A.2.d), and standardised (Section A.2.g) assay that, through accrual of validation data, provides less biased and more precise estimates of DSe and DSp. These estimates, along with evidence-based data on prevalence of infection in the population being tested, are the basis for providing a high degree of confidence in the predictive values of positive and negative test results. In order to insure that test results provide useful diagnostic inferences about animal population infection/exposure status, the validation process encompasses initial development and assay performance documentation, as well as ongoing assessment of quality control and quality assurance programmes. Figure 1 shows the assay validation process, from assay design through the development and validation pathways to implementation, deployment, and maintenance of the assay

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.

a) Fitness for purpose

The most common purposes are to:

1) Demonstrate 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) Eradication of disease or elimination of infection from defined populations.

4) Confirmatory diagnosis of suspect or clinical cases (includes confirmation of positive screening test).

2 This is a specific interpretation of the more generally stated requirements of the ISO/IEC 17025:2005 international quality standard for testing laboratories (18). 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 *Terrestrial & Aquatic Manuals.*

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 Appendices 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.

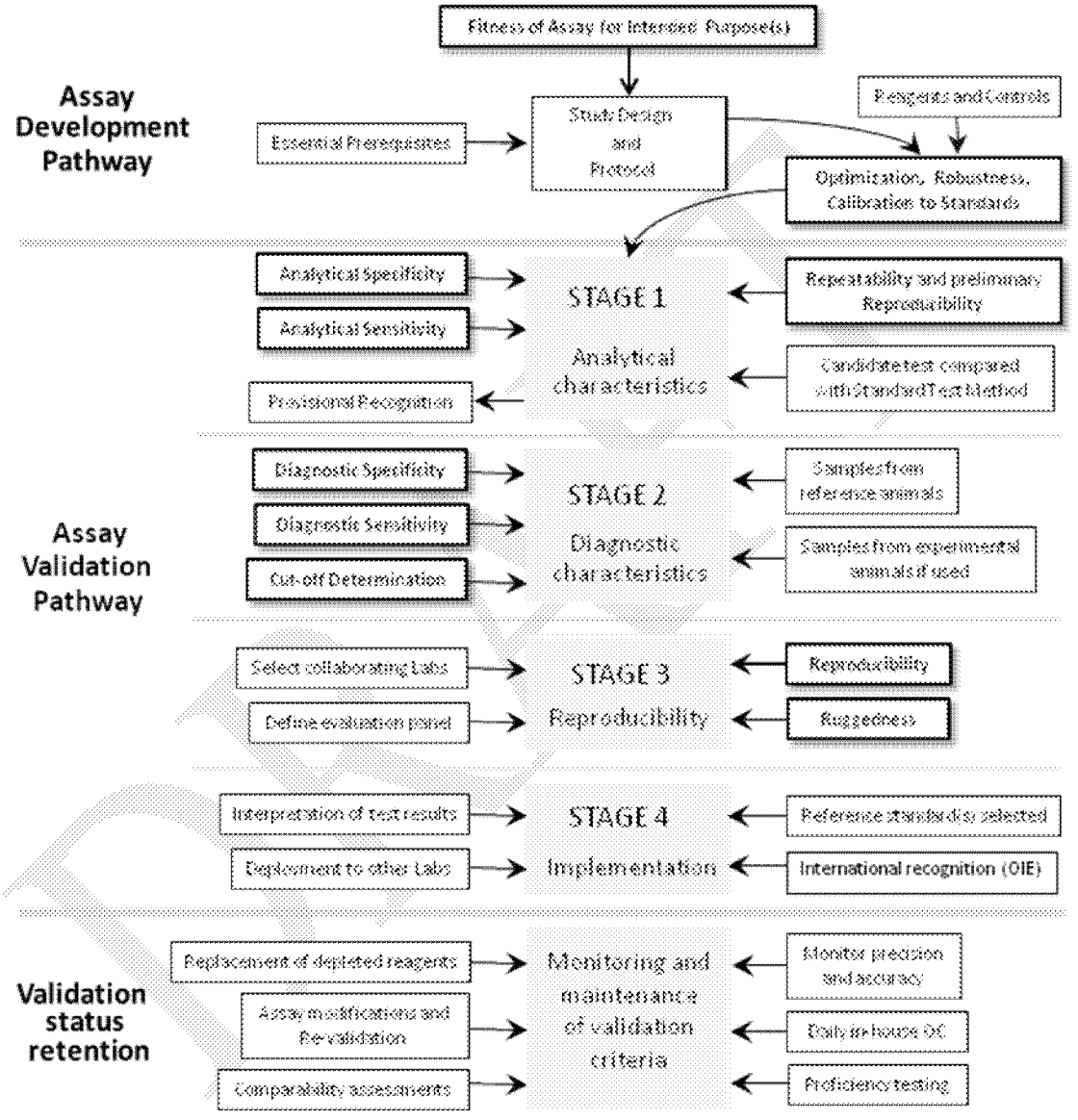

Figure 1. *The assay development and validation pathways with assay validation criteria highlighted in bold typescript within shadowed boxes.*

b) 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 relevance 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. 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 (8).

2. Assay development – the experimental studies

a) Essential prerequisites: factors that impact assay validation

i) **Quality assurance.** Whether developing assays in the laboratory or performing analyses of clinical material, the objective is to produce data of high quality. This requires that key requirements have to be fulfilled within the laboratory (see Chapters 1.1.3 & 1.1.1 of the *Terrestrial & Aquatic Manuals*, respectively) The establishment of quality assurance (QA) and quality control (QC) systems is essential, i.e. a set of quality protocols, including the use of assay control samples that ensure that the system is working properly and confirms data reproducibility and quality. QA and QC systems, together with trained and competent personnel, have already been established in many laboratories world-wide.

ii) **Equipment selection.** Equipment that is not maintained and calibrated can be a major impediment to achieving a quality assay. Apparatus (freezers, heating blocks, incubators, refrigerators, optical colorimeters, thermocyclers, plate washers, pipettes, etc.) must be calibrated according to the laboratory's quality assurance protocols. Examples of this need include robotics used for automation of entire assays, or parts thereof, for routine diagnostic processing. It is not sufficient to assume that robotic extraction of nucleic acid, for example, is equivalent to previously used manual extraction methods or that an automated ELISA plate washer provides uniform washing among wells of the plate. The instrument must be calibrated and the protocol validated to confirm performance efficiency and to assure cross-contamination does not occur in NAD systems or washing is adequate for all wells in a plate. (See Appendices on best practices for more details).

iii) **Selection and integrity of samples.** Selection, collection, preparation and management of samples are critical variables in designing, developing, and validating an assay. Other variables such as transport, chain of custody, tracking of samples, and the laboratory information management system are also critical 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 in experimentation or routine diagnostic testing. 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 used in the assay (e.g. serum, tissue, whole blood) and representative of the species to be tested by the resulting assay. The reference materials should appropriately represent the range of analyte concentration to be detected by the assay. Details on proper sample collection, preparation, management, and transport are available in the Chapter 1.1.1 of the *Terrestrial Manual*.

b) Test method design and proof of concept

Considerable thought and planning needs 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 Appendices to this chapter, which cover best practices for development and validation of assays for detection of various analytes (e.g. antibody, antigen, and nucleic acid detection).

i) **Analyte reference samples**. Development of all assays is dependent on analyte reference samples that reflect the target analyte and the matrix in which the analyte is found in the population for which the assay is intended. The reference samples may be sera, fluids 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.

Analyte reference samples, containing the analyte of interest in varying concentrations, are useful in developing and evaluating the candidate assay's validation criteria.

c) Operating range of the assay

During development of the assay, the lower and upper detection limits are established. To formally establish 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 in an analyte-negative matrix of the same constitution as the sample matrix of samples from animals in the population targeted by the assay. The results are plotted

Operating range of an assay: an interval of analyte concentrations (amounts) over which the method provides suitable accuracy and precision.

as a 'response-curve', with the response (e.g., OD, Ct, etc) a function of analyte concentration (amount). The curve, establishes the range of the assay, which is the interval between the upper and lower concentration (amounts) of analyte in the sample for which a suitable level of precision[3] and accuracy has been demonstrated. For most diagnostic assays, the response is the result of interaction of the analyte with an antibody or other binding reagent. These are known as ligand binding assays (LBAs), The typical calibration curve for LBAs is sigmoidal in shape, with a lower boundary (asymptote) near the background response (non-specific binding) and an upper asymptote near the maximum response, Typically, LBA data are transformed to approximate a linear relationship between response and concentration. This transformation simplifies interpolation of data using linear regression analysis, but with the disadvantage of introduced bias. Since linearization is imperfect leading to compromised estimates of accuracy and precision, numerous data-fitting algorithms have been applied to experimental calibration curve data from LBAs (11). The currently accepted reference model for calibration of LBAs is the 4-parameter logistic model, which usually optimizes accuracy and precision over the maximum usable calibration range (11). Such transformations are now more practical for the general user because of many user-friendly statistical software programs available on the internet.

Precision is the degree of dispersion among a series of measurements of the same sample tested under specified conditions (see footnote 3 for more detail)

Accuracy is the closeness of of a test value to the expected (true) value for a reference standard reagent of known concentration or titer

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.

d) Optimisation

It is useful to select at least three well-defined reference samples, representing the analyte ranging from high positive to negative (e.g., negative, low- and high-positive). 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. 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 because the matrix of field samples may be very different from spiked-sample matrix. But, when no other alternative exists, spiking a sample with a known amount of the analyte 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 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, 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 quality assay. Scientific judgment and use of best scientific practices provided in accompanying Appendices to this chapter are the best assets to guide optimisation of all elements of an assay. The approach outlined provides a solid foundation for development of an assay that fulfils the criteria of "robustness" and "ruggedness" when used over extended periods of time within a laboratory or when implemented in other laboratories. 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, optimization becomes extremely 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 for parameters such as pH, molarity, etc. Likewise for biologicals, standards for quality, purity, concentration and reactivity must be defined. Shelf lives and storage conditions must also be considered for both chemicals and biologicals. Acceptable ranges for reaction times and temperatures need also be established. Essential equipment critical to assay performance must be described in detail, including operational specifications and calibration. Process (quality) control is often an add-on at the end of assay development but it should be an integral part of optimization from the very beginning. In addition to the above, downstream aspects such as

[3] Precision may be evaluated in several ways by testing the same replicated sample: 1) within a plate or plates in a run of the assay, 2) between plates run concurrently within a run of the assay, 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, precision categories 1-3 are estimates of repeatability, and precision category 4 is synonymous with reproducibility. Levels 3a and 3b are also known as intermediate precision.

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

In some assay types, a correct assay result is fully dependent on getting a particular step in the testing process correct, requiring special attention during optimisation. A case in point is nucleic acid extraction from the sample. Both commercial (robotic, spin columns, and magnet-based extractions, etc.) and standard chemistry-based methods are used for DNA or RNA extraction. It is crucial to determine the most reproducible and efficient extraction method through optimisation experiments. Extraction needs to be optimised for every type of tissue that may be targeted by the assay. If the method of extraction is changed, at a minimum, comparable efficiency of extraction should be demonstrated (see Section B.6 below and associated Appendix 1.1.4.6 for additional information on establishing comparability when reagents 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, proper preparation and storage of all biological reagents and reference materials must be heeded to ensure stability (see Chapter 1.1.1 of the *Terrestrial Manual*).

During experimentation to optimise the assay, take note of assay parameters that have a narrow range in which they perform optimally, as these are the critical points that may affect an assay's robustness (see Section A.2.f).

e) Inhibitory factors in sample matrix

Generally, for antibody detection in serum, assays are rather resistant to inhibitory factors, with the exception of certain assays, e.g. toxic factors in viral neutralisation assays, or when endogenous substances found in certain sample types inhibit enzymatic reactions in ELISAs. For nucleic acid detection, sample matrices including blood, serum, body tissues, and swab samples allow for easy extraction of target nucleic acids, while faeces, autolysed tissues and semen samples can be more difficult to handle because of the presence of factors which can inhibit downstream assays such as PCR.

f) Robustness

Robustness is a measure of an assay's capacity to remain unaffected by small, but deliberate, variations in method parameters, and provides an indication of its reliability during normal usage.

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. It is assessed by deliberate variations in method parameters (23). Assessment of robustness should begin during assay development and optimization 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 optimizing 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 exhibit ruggedness (reduced reproducibility) when transferred to other laboratories (ruggedness is addressed in Section 4).

The factors most likely to affect assay robustness are quantitative (continuous) such as pH and temperature; qualitative (categorical) such as batch of reagents or brand of microtiter plates; and mixture-related such as aqueous or organic matrix factors (9). For ligand-binding assays (LBAs), lack of robustness is not only due to less-than-optimal concentration/amount of the bio-reagent specified in the method, but may also be due to the intrinsic characteristics of the biological reagent (e.g. monoclonal antibody affinity or polyclonal antibody avidity and/or valency). Robustness, therefore, particularly of LBA-based assays, may be affected by systematic and/or random errors (22).

Robustness testing is demonstrated on a method-by-method basis. All critical reagents are identified and subjected to a factorial design experiments which compares all possible combinations of reagents (9, 26). For example, in antibody detection by ELISA, factors may include concentration of antigen bound to the solid phase, conjugate dilution and several test sera representing the operating range of assay. The response of the assay with respect to these small changes shall not result in unacceptable variability.

Robustness is further verified during Stage 1 of assay validation. When the optimized test is first run under routine laboratory conditions, this practical measure of robustness is referred to as repeatability (see Section 2.a) and it is continually monitored as part of process control procedures for the duration of the life of the assay (see Section 6.a).

g) Calibration of the assay to standard reagents

i) **International and national analyte reference standards.** Ideally, international reference standards, containing a known concentration of analyte, are the reagents to which all assays are standardised. Such standards are prepared and distributed by international reference laboratories. National reference standards are calibrated by comparison with an international standard reagent 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 standard reagents are highly characterised through extensive analysis, and preferably the methods for their characterisation, preparation, and storage have been published in peer-reviewed publications.

ii) **In-house standard reagent.** An in-house reference standard generally has the highest metrological quality available at a given location in a given organization, and is 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 analyte standards. 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.

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

h) "Normalising" test results to a working standard(s)

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 virtually impossible to directly compare (semi-) quantitative data. To markedly improve 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' [not to be confused with transformation of data to achieve a 'normal' (Gaussian) distribution]. The 'normalized' values may be expressed in many ways, such as a percent of a positive control (e.g. in an ELISA), or as a concentration or titer of an analyte derived from a standard curve, or as a number of targeted genomic copies also derived from a standard curve of Ct (cycle threshold) values for real time PCR. It is good practice to include working standards [or at least a reasonably well-characterized sample(s) in all runs of the assay during assay development and validation because this allows 'normalization' of data which provides a valid means for direct comparison of results between runs of an assay. Automated assay systems may calculate and report 'normalized' data by, for example, a standard curve, or by reporting the cycle number at which the cycle threshold is exceeded as in real-time PCR. For more information, see Appendices 1.1.4.1 and 1.1.4.3.

B. ASSAY VALIDATION PATHWAY

1. Definition of a validated assay.

"Validation" is a process that determines the fitness of an assay that has been properly developed, optimised and standardised for an intended purpose. 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)"

To retain the status of a validated assay, however, it is necessary to assure that the assay as originally validated consistently maintains the performance characteristics as defined during validation of the assay (see Section 6 below). This can be determined in a quality assurance program characterized by carefully monitoring the assay's daily performance, primarily through precision and accuracy estimates for internal controls, and by scheduled external proficiency testing. Should the assay cease to produce results consistent with the original validation data, the assay would be rendered unfit for its intended purpose. Thus, a validated assay must be continuously assessed to assure it maintains its fitness for purpose.

2. Stage 1 – Analytical performance characteristics

Ideally, the design for studies outlined in the following sections should be done with assistance of a statistician. It is possible to design experiments that efficiently provide information on likely within- and among-laboratory sources of variation in assay precision (see footnote 3 under Section A.2.c, above) which will define the

performance characteristics of the assay. Stage 1 studies for repeatability, reproducibility and assessment of analytical sensitivity (limit of detection) should be performed in a blinded fashion with random selection of samples. The choice of organisms, strains or serotypes to assess analytical specificity should reflect current knowledge and therefore inform the best possible experimental design for targeting specific analytes.

a) 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 from a minimum of three (preferably five) samples representing analyte activity within the operating range of the assay. Each of these samples is then aliquoted into individual vessels as three identical replicates of the original sample containing the original analyte and matrix concentration. 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 (approximately 20) involving two or more operators, done on at least five separate days. The variation in replicate results can be expressed as standard deviations, confidence intervals, or other possible options (see Appendix 1.1.4.4 on Measurement Uncertainty for assessments of repeatability).

b) Analytical specificity (ASp)

Analytical specificity is the degree to which the assay distinguishes between the target analyte and other components that may be detected in the assay.

Analytical specificity distinguishes between the target analyte and other components in the assay in at least three distinctive ways. These are described as the selectivity, exclusivity, and inclusivity of the assay.

- **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) non-specific binding of reactants to a solid phase, (e.g. conjugate of an ELISA adsorbed to well of microtiter plate); 4) antibodies to vaccination which maybe 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 (24) is a useful overview of selectivity as defined for analytical chemistry from which a modification described herein was deduced for application to diagnostic tests.

- **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.

- **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 characterizes the scope of action for a screening assay.

After eliminating interferents to the extent possible, the next step is to test a panel of samples appropriate for evaluating either inclusivity, exclusivity, or both, depending on the intended purpose of the assay.

ASp applies to both direct and indirect methods of analyte detection. If exclusivity is required, field samples should be obtained from animals infected with genetically-related, non-pathogenic organisms, but this may prove difficult or even impossible. In such cases, cultivated organisms can be used in direct detection methods, or serum from animals exposed experimentally by natural routes for indirect detection methods. Acceptable cross-reactivity is largely dependent on the intended purpose of the test and the prevalence of the cross-reactive organisms/analytes in the target population samples – which must be determined for each case (see Appendix 1.1.4.5 for more detail). For PCR, it is useful to perform computer simulation studies as an adjunct to laboratory assessment of ASp; however, such studies are not sufficient by themselves to evaluate ASp.

A factor unique to viral antibody detection is the possible antibody response of animals to carrier proteins found in vaccines – another type of interferent that may negatively affect selectivity. If such proteins are also present in the solid phase antigen on ELISA plates, they may bind antibody developed to vaccine carrier proteins and give false-positive results (lack of exclusivity in the assay). Use of vaccine preparations as antigens in ELISAs is, therefore, not recommended. See Appendix 1.1.4.1 for specific practices to determine ASp.

c) Analytical sensitivity (ASe)

Analytical sensitivity is synonymous with the lower limit of detection (LOD) of an analyte in an assay. For direct-detection assays, this may be expressed as the number of genome copies, infectious dose, colony-forming units, plaque forming units, etc., of the agent that can be detected and distinguished from the result of a matrix background. Most commonly, this is expressed as a number of copies, complement-fixing units or plaque-forming units giving at least 50% positive results among the replicates of a sample for a specified volume or weight (see Appendix 1.1.4.5 for detail on LOD determination). For indirect detection assays, it is the least amount of antibody detected, usually, the penultimate dilution of sample in which the analyte is indistinguishable from the activity of a sample matrix control.

Limit of detection or analytical sensitivity, is the smallest amount of analyte in a sample that can be detected by a direct detection assay in at least 50% of the replicates for each dilution, in a dilution series of analyte in matrix.

If the intended purpose is to detect low levels of analyte or subclinical infections and it is difficult to obtain the appropriate reference materials, for example samples from early stages of the infection process, it may be useful to determine comparative analytical sensitivity by running a panel of samples on the candidate assay and on another independent assay. This would provide a relative comparison of analytical sensitivity between the assays, but care must be taken in choosing the independent assay used in the comparison to ensure that the analytes being detected (if different) demonstrate the same type of pathogenic profile in terms of time of appearance after exposure to the infectious agent, and relative abundance in the test samples chosen.

Where a new, more sensitive test is developed, it may be necessary to test serial samples taken from infected animals early after infection, on through to the development clinical or fulminating disease, and to run these in parallel with previously used tests to demonstrate the increased sensitivity. This would also provide a temporal comparison of the earliest point of detection relative to the pathogenesis of the disease.

d) Standard test method for comparison with the candidate assay test method

There are situations where it is not possible or desirable to continue to Stage 2 of the Validation Pathway because appropriate samples from the target population are scarce and animals are difficult to access (such as for exotic diseases). However, a small but select panel of highly characterised test samples representing the range of analyte concentration should be run in parallel on the candidate assay method and by the standard method if it exists.

A **Standard test method** is the best internationally or nationally recognized method, or in their absence, the best available method preferably published in a reputable journal.

e) 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, or molecular sequencing to confirm a real-time PCR test result. Pathogenicity indices, haemagglutination inhibition, drug resistance determinations, etc. are other examples where adjunct tests or procedures are performed, either independent of, or as part of a primary assay.

Such adjunct tests must be validated for analytical performance characteristics (Sections A.2 through B.2.c., above), but differ from routine diagnostic tests because they do not require validation for diagnostic performance characteristics (Sections B.3 through B.5, below). 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.

f) Preliminary evaluation of reproducibility

Preliminary reproducibility estimates of the candidate assay may be useful at this time in the validation process, where only a small panel of highly characterised samples is available. This panel could be used for a limited evaluation of reproducibility to enhance provisional acceptance status for the assay. The candidate test method is then duplicated in laboratories at one or more 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 Stage 3 of assay validation.

g) Provisional assay recognition[4]

Provisional recognition defines an assay that has been assessed through Stage 1 for critical assay benchmark parameters: ASe, ASp, repeatability, and an estimate of reproducibility.

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 (see requirements for Stage 2, Diagnostic Performance, below). 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 amount of allowable error and the level of confidence in the estimates (Table 1 and reference 19). 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 threshold or if no adjustment is needed, enhance confidence in the estimates.

Historical precedent would suggest that assays were generally the product of laboratory experiments with an emphasis on analytical sensitivity and analytical specificity, and evaluation of panels of field samples was nominal. Such bench validation for bovine spongiform encephalopathy (BSE) is a classical example where positive field samples were not available. Nevertheless, during extended periods of usage in diagnostic settings, such tests have undergone adjustments based on empirical evidence to reduce false-positive and false-negative results. For some of the BSE rapid tests, adjustments had to be made in the cut-off to reduce false-positive results apparent in early implementation. Bench validation provided a level of confidence in diagnostic performance that was adequate for conditional diagnostic use as determined by national authorities. But, it must never become a replacement for a full field validation. Therefore bench validation of diagnostic assays can only offer provisional recognition with anticipation that full field validation will follow.

A provisional recognition of an assay by state or national authorities recognises 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. This concept should be limited to emergency situations where rapid introduction of new tests is deemed essential by authorities. There may be other situations where bi-lateral trade agreements, based on tests (e.g. standard test methods) that have not been fully validated within a given country, are mutually accepted. In exceptional cases for rare diseases where no other assay option exists, provisional recognition may be allowed, but reporting of results should include a statement of the provisional nature of the validation of the assay. In all cases, sound evidence must exist for preliminary estimates of DSp and DSe based on a small select panel of well-characterised samples containing the targeted analyte.

3. Stage 2 – Diagnostic performance of the assay

Diagnostic Sensitivity and Diagnostic Specificity. Estimates of DSe and DSp 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. Receiver operating characteristic curve analysis is a useful adjunct to estimation of DSe and DSp because it assesses the global accuracy of a quantitative diagnostic test across possible assay values (14, 15, 27). This approach is described in-depth in Appendix 1.1.4.5.

Diagnostic sensitivity is the proportion of samples from known infected reference animals that test positive in an assay.

Diagnostic specificity is the proportion of samples from known uninfected reference animals that test negative in an assay.

A sampling design must be chosen that will allow estimation of DSe and DSp. 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 reference 19). An abbreviated 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

[4] Provisional recognition does not imply certification 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, usually for a time-limited use in emergency situations or as the basis for bi-lateral agreements between countries that choose to accept results from such an assay for trade purposes.

DSe or DSp. Comparison of a 5% vs 2% error shows a considerable reduction in the number of samples required. A rather large number of samples is required to achieve a very high confidence for DSe and DSp when a minimal amount of error in the estimate is desired. Logistical and financial limitations may require that less than an optimal number of samples will be evaluated. However, by reducing the DSe and DSp confidence levels to less than 90% usually would not be recommended. Sample size also may limited by the fact that reference populations and "gold standards" (perfect reference standards) may be lacking (see Appendix 1.1.4.5 for further detail). It may, therefore, be necessary to use a sub-optimal number of samples initially. It is, however, highly desirable to enhance confidence and reduce allowable error in the DSe and DSp estimates by adding more samples as they become available.

Table 1. Theoretical number of samples from animals of known infection status required for establishing diagnostic sensitivity (DSe) and specificity (DSp) estimates with known 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					
	75%	80%	85%	90%	95%	99%	75%	80%	85%	90%	95%	99%
90%	257	369	475	610	864	1493	41	59	76	98	138	239
92%	210	302	389	466	707	1221	34	48	62	75	113	195
94%	161	232	298	382	542	935	26	37	48	61	87	150
95%	136	196	251	372	456	788	22	31	40	60	73	126
96%	110	158	203	260	369	637	18	25	32	42	59	102
97%	83	119	154	197	279	483	13	19	25	32	45	77
98%	56	80	103	133	188	325	9	13	16	21	30	52
99%	28	41	52	67	95	164	4	7	8	11	15	26

- Percent error allowed in the estimate of DSe or DSp = 2% in the left panel and 5% in the right panel. For the number of samples required for 1%, 3%, and 4% allowable error in the estimate of DSe and DSp, multiply the number of samples in the left panel of the table by a factor of 4.0, 0.44, and 0.25, respectively.

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

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.

i) **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 that have eradicated or have never had the disease in question. Such samples are useful as long as the targeted population for the assay is similar to the sample-source population.

ii) **Positive reference samples.** It is generally problematic to find sufficient numbers of true positive reference animals, as determined by isolation of the organism. It may be necessary to resort to samples from animals that have been tested by another test such as a nucleic acid detection system.

iii) **Samples from animals of unknown status.** See Appendix 1.1.4.5 Section 5.b.i, and Section 3.b.ii of this chapter for a discussion of latent class models.

b) Reference animal infection status

i) **So-called "Gold Standard" model.** The term 'gold standard' is commonly used to describe any standard of comparison, but it should be limited to methods or combination of methods that unequivocally classify animals as infected/exposed or uninfected. Some isolation methods themselves have problems of repeatability and analytical sensitivity, so are not truly gold standards particularly for purportedly negative samples. When the so-called reference standard is imperfect, which is the common scenario for most ante-mortem tests, estimates of DSe and DSp for the candidate assay may be compromised because the error in estimates obtained by comparison to the relative standard is carried over into the estimates for the candidate assay. Indeed, when using imperfect reference assays, the DSe and DSp performance estimates of the candidate assay will be flawed and often overestimated.

NAD assays may be more sensitive and specific than existing 'gold standard' methods, which render the established 'gold standard' as not suitable for use as a comparison. If the NAD is more sensitive than the 'gold standard,' an apparent lower relative specificity will be misleading. This problem may be partially resolved by assessing sample derivation, clinical history and sequencing of any PCR products to confirm analyte identity.

ii) Latent-class models. Latent-class models (5, 10, 12 17) 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 . 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 Appendix 1.1.4.5 on statistical considerations for details).

c) Experimentally infected or vaccinated reference animals

Sera 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. Only single time-point sampling of individual experimental animals is acceptable. Also, for indirect methods of analyte detection, exposure to organisms under experimental conditions, or vaccination may elicit antibody responses that may not be quantitatively and qualitatively typical of natural infection in the target population (19). 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 samples from experimental animals.

d) Threshold (cut-off) determination

To obtain DSe and DSp estimates of the candidate assay, the test results first must be reduced to categorical (positive, negative, or intermediate) status. This is accomplished by insertion of one or two cut-off points (threshold or decision limits) on the continuous scale of test results. The selection of the cutoff(s) should reflect the 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 (5, 12, 13, 15, 19, 27 and Appendix 1.1.4.5 on statistical considerations). If considerable overlap occurs in the distributions of test values from known infected and uninfected animals, it is difficult 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). The values that fall between these percentiles may be classified as intermediate (see box), and would require testing by a confirmatory assay, retesting for detection of sero-conversion, or sequencing for identity.

Threshold and cut-off are considered to be synonymous. A cut-off is the test value selected for distinguishing between negative and positive results on a continuous scale of test values.

Intermediate, inconclusive, suspicious, or equivocal are terms used synonymously for a zone of test values between the positive and negative cut-offs

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.g. on provisional acceptance of an assay during accrual of data to enhance estimates of DSe and DSp.

e) Calculation of DSe and DSp based on test results of reference sera

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 X 2 table. In a hypothetical example, assume the test developer decided that estimated DSe and DSp for the new assay should be 97% and 99%, respectively, with a desired confidence of 95% for both estimates. The amount of allowable error 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 an hypothetical set of results and the calculated DSe and DSp estimates based on the samples tested.

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)	7 (FP)
	Negative	9 (FN)	88 (TN)
		Diagnostic sensitivity* TP/(TP + FN) **96.7%** (94.0 – 98.5%)**	Diagnostic specificity* TN/(TN + FP) **92.0%** (84.3 – 96.7%)**

* 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% = Allowable error 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 Appendix 1.1.4.5 for information on confidence limits)

In this example, the DSe estimates are as anticipated, but the DSp is much reduced from the anticipated 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.

4. Stage 3 – Reproducibility and augmented repeatability estimates

Reproducibility is an important measure of the precision of an assay when used in several laboratories located in distinct or different regions or countries using the identical assay (protocol, reagents and controls). Each of at least three laboratories test the same panel of samples (blinded) containing a minimum of 20 samples, with identical aliquots going to each laboratory (see Appendix 1.1.4.7 on panels of samples). This exercise also generates preliminary data on non-random effects attributable to deployment of the assay to other laboratories - also known as ruggedness of the assay. 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 Appendix 1.1.4.4 on Measurement of Uncertainty for further explanation of the topic and its application).

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.

Ruggedness is a measure of the assay's capacity to remain unaffected by substantial changes or substitutions in test conditions anticipated in multi-laboratory utilisation.

5. Stage 4 – Programme implementation

a) Interpretation of test results.

Predictive values of test results. An assay's test results are most useful when the inferences made from them are accurate. Predictive values for test results need to be based on the true prevalence of exposure/infection in the targeted population. For screening assays used in surveillance of a 'disease-free' population, false-positive results are a significant problem. For instance, an assay may have impeccable credentials (e.g. high precision and accuracy, 99% DSe and 99.9% DSp). But if the prevalence of disease is close to zero, and the assay

Predictive Value (PV) of positive or negative test result. PV+ is the probability that an animal has been exposed or infected, given that it tests positive. PV is the probability that an animal is free from exposure or infection, given that it tests negative.

has one false positive for every 1000 animals tested, false-positive inferences are a problem (see reference 19 for PV tables for various estimates of DSe and DSp). Similarly, if the assay for a highly virulent disease has one false negative for every 100 animals, false negative inferences could have devastating consequences. This illustrates the critical importance of choosing diagnostic thresholds that are appropriate for the application at hand. Thresholds should be chosen to minimize the effect of false positives and/or false negatives on the predictive values of the test given its application and the prevalence of exposure/infection in the target population. It may also be prudent to have highly specific confirmatory assays to determine whether screening assay reactors are true or false positives.

For nucleic acid assays, it may be necessary to confirm NAD-positive results by sequence analysis of the amplified product (an example of an assay to assist in resolving errors due to non-specific target or primer binding).

b) 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 test kits that have completed the certification process, 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 test users with an informed and unbiased source of information about the test and its performance characteristics for an intended purpose. The Register is available on the OIE website at: www.oie.int/vcda/eng/en_vcda_registre.htm

c) 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 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 ruggedness should be repeated when the test is transferred from the development laboratory to the field, whether for use in local laboratories or in pen-side 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 ruggedness (which is mostly derived from reproducibility estimates).

6. Monitoring assay performance after initial validation

a) Monitoring the assay

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.

A validated assay in routine use needs to be consistently monitored for repeatability through process controls to evaluate possible temporal changes in test precision and accuracy. These changes can be monitored graphically by plotting control values in control charts. 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. Subsequent ongoing evaluation of the assay's performance is also essential and is usually done through assessments of precision, accuracy, and outlier tendencies using control charts. Reproducibility is assessed through external quality control programmes such as proficiency testing.

b) Modifications and enhancements - considerations for changes in the assay

Over time, modification of the assay likely will be necessary to address changes in the analytes targeted (i.e., modification of the assay to adjust diagnostic performance) or technical modifications may be needed to improve or enhance assay efficiency or cost-effectiveness.

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 a virus, derived from animals in different geographic locations, are known to have different target sequences or primer sites, requiring revalidation of the assay. This is especially true for NAD systems as it is very common for point mutations occur in many

infectious agents (i.e. 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 estimates of DSe for the assay are not compromised.

A similar situation may occur with incursion of new viral lineages into countries or regions where that viral lineage did not previously exist. In these circumstances, existing NAD assays which did not target these novel lineages may need to be modified to include primers or probes targeting these new analytes. The same would be true for typing sera used in virus neutralisation assays.

i) **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 modifications of the assay affect the test results. 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 Appendix 1.1.4.6 for description of experiments that are appropriate for comparability testing and Appendix 1.1.4.7 on reference sample panels.

ii) **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 in an NAD system). It may include changes to reagents themselves (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 or enhancement 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 would require 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 Appendix 1.1.4.6 for a description of comparability assessment using diagnostic samples. If the comparability assessment does not suggest any significant change in diagnostic performance, the modified assay may be phased into routine use. If, on the other hand, significant differences are observed, the modified assay would require additional Stage 2 or field validation before being adopted.

iii) **Replacement of depleted reagents.** When a reagent such as a 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 should be included in multiple runs of the assay in parallel with the original control to establish their proportional relationship. Whenever possible, it is important to change only one reagent at a time to avoid the compound problem of evaluating more than one variable.

c) 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 from animals of known infection status. This 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.

REFERENCES

1. BALLAGI-PORDÁNY A. & BELÁK S. (1996). The use of mimics as internal standards to avoid false negatives in diagnostic PCR. *Mol. Cell. Probes,* **10**, 159–164.

2. BELÁK S. & THORÉN P. (2001). Molecular diagnosis of animal diseases. *Expert Rev. Mol. Diagn.*, **1**, 434–444.

3. BELÁK S. (2005). The molecular diagnosis of porcine viral diseases: a review. *Acta Vet. Hung.*, **53**, 113–124. (Review).

4. BELÁK S. (2007). Molecular diagnosis of viral diseases, present trends and future aspects. A view from the OIE Collaborating Centre for the Application of Polymerase Chain Reaction Methods for Diagnosis of Viral Diseases in Veterinary Medicine. *Vaccine*, **25**, 5444-5452.

5. BRANSCUM A.J, GARDNER I.A, JOHNSON. W.O. (2005). Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Preventive Veterinary Medicine* 68, 145–163.

6. BURNS M.J., NIXON G.J., FOY C.A. & HARRIS N. (2005). Standardisation of data from real-time quantitative PCR methods - evaluation of outliers and comparison of calibration curves. *BMC Biotechnol.*, **5**, 31–44.

7. BUSTIN S.A. (2005). Real-time, fluorescence-based quantitative PCR: a snapshot of current procedures and preferences. *Expert Rev. Mol. Diagn.*, **5**, 493–498..

8. 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.

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

10. 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.

11. FINDLAY J.W.A, DILLARD R.F. *(2007)* . Appropriate Calibration Curve Fitting in Ligand Binding Assays. *AAPS Journal.* **9**(2): E260-E267. (Also on-line as AAPS Journal **(**2007); **9** (2), Article 29 (http://www.aapsj.org)

12. 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, pp. 63–76.

13. 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.

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

15. 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.

16. HUGGETT J., DHEDA K., BUSTIN S. & ZUMLA A. (2005). Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun.*, **6**, 279–284. (Review).

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

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

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

20. LAUERMAN L.H. (2004). Advances in PCR technology. *Anim. Health Res. Rev.*, **5**, 247–248.

21. LOUIE M., LOUIE L. & SIMOR A.E. (2000). The role of DNA amplification technology in the diagnosis of infectious diseases. *CMAJ.,* **163**, 301–309.22.

22. THOMPSON, M., ELLISON, S. & WOOD, R. (2002). Harmonized guidelines for single-laboratory validation of methods of analysis. *Pure Appl. Chem.*, **75** (5), 835-855.

23. VALIDATION OF ANALYTICAL PROCEDURES: TEXT AND METHODOLOGY (Q2(R1). *(2005).* 1-13 Tripartite Guideline of the International Conference on Harmonisation for Technical Requirements for Registration of Pharmaceuticals for Human Use. Pp 1-13 (Available on-line at: http://www.ich.org/LOB/media/MEDIA417.pdf

24. 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.

25. 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.

26. YOUDEN, W. & STEINER, E. (1987). Statistical Manual of the AOAC. AOAC International. 96 Pages. 5th Printing (ISBN 0-935584-15-3).

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

*
* *

APPENDICES TO VALIDATION CHAPTER (NOTE: Titles may change)

Appendix 1.1.4.1. Development and optimisation of antibody detection assays

Appendix 1.1.4.2. Development and optimisation of antigen detection assays by immunological means

Appendix 1.1.4.3. Development and optimisation of Nucleic Acid Detection (NAD) assays

Appendix 1.1.4.4. Measurement Uncertainty

Appendix 1.1.4.5. Statistical Approaches to Validation

Appendix 1.1.4.6. Comparability of Assays after Minor Changes in a Validated Test Method

Appendix 1.1.4.7. Reference Samples and Panels – Selection and Use