CHAPTER 1.1.6.

PRINCIPLES AND METHODS OF VALIDATION OF DIAGNOSTIC ASSAYS FOR INFECTIOUS DISEASES

INTRODUCTION

Adequate validation and verification of the performance characteristics of diagnostic tests for infectious diseases are critical to ensuring that assays are applied and interpreted in a scientifically robust and defensible manner (Colling & Gardner, 2021). Since it was first adopted in 1996, the World Organisation for Animal Health (WOAH: founded as the OIE) Assay Development and Validation Pathway (Figure 1) has acted as the internationally recognised standard for the validation of veterinary diagnostic tests for infectious diseases.

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 regardless of whether they are used in the laboratory or as point-of-care tests should be validated for the species and specimen 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, the assay's performance should be carefully monitored under conditions of routine use, often by tracking the behaviour of assay controls within each run and through on-going assessment during routine diagnostic use in the targeted population over time. Should it no longer produce results consistent with the original validation data, the assay may be rendered unfit for its intended purpose(s).

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 (see Section A.1 Definition of the intended purpose(s) of an assay*).*

This chapter focuses on the criteria that must be fulfilled during assay development and validation of all assay types and the metrics used to characterise test performance. 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. The guiding principles described herein also apply to infectious diseases that are not WOAH listed. More detailed guidance is provided in a series of recommendations for validation of diagnostic tests (Section 2.2 Validation of diagnostic tests, *chapters 2.2.1 to 2.2.8) that are tailored for several fundamentally different types*

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

of assay (e.g. detection of antibodies, antigen and nucleic acids) and provide more information on specific issues related to the validation of diagnostic assays (Halpin et al., *2021). For specific information for wildlife species, refer to Chapter 2.2.7* Principles and methods for the validation of diagnostic tests for infectious diseases applicable to wildlife (*Jia* et al., *2020; Michel* et al., *2021*). *The information provided in chapter 2.2.7, 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.*

An up-to-date compilation of the relevant validation standards (WOAH and non-WOAH) and guidance documents for all stages of diagnostic test validation and proficiency testing, including design, analysis, interpretation as well as clear reporting and case studies are provided in the WOAH Scientific and Technical Review *issue Diagnostic Test Validation Science (Vol. 40, April 2021). Published standards for peer-reviewed reporting of accuracy studies (STARD) are available for infectious diseases of human, terrestrial (paratuberculosis, Bayesian latent class model) (Bossuyt* et al., *2015; Gardner* et al., *2011; Kostoulas* et al., *2017) and aquatic animals (Gardner* et al., *2019; Kostoulas* et al., *2021). Verification (Kirkland & Newberry, 2021) and comparability studies (Reising* et al., *2021) are briefly described at the end of this chapter for assays that have completed at least stage 2 of the WOAH pathway. There is a pressing need to develop validation guidelines and standards for the rapidly increasing use of point-of-care tests (POCTs). Typically, POCTs are used in the field under varying environmental conditions, on a range of sample types collected in non-sterile settings, by operators with a diverse range of experience, training and proficiency. Field-testing conditions, including extreme variations in temperature and humidity, as well as other variables such as water and reagent quality, an inadequate cold chain, operator ability, and poor or non-existent quality assurance systems, can all contribute to lower test accuracies than those reported by POCT manufacturers or obtained in an accredited laboratory (see Figure 1 Stage 5). The consequences of a positive test result and the need for confirmatory testing and reporting by an accredited laboratory in particular when performing a test for an exotic disease need to be embedded in existing testing policies and guidelines. POCT-specific standards and recommendations, such as the point-of-care key evidence tool (POCKET) checklist for multi-dimensional evidence reporting; scorecards and guidelines for POCT evaluation; and guidelines on quality practices in non-instrumented POCTs (i.e. those that do not require a specific piece of equipment) and ISO/TS 22583:2019 Guidance for supervisors and operators for POCT devices provide guidance for healthcare workers. In some countries, such as Germany, POCTs that detect notifiable and reportable animal diseases require formal authorisation by the national licensing authority. In other countries, POCT accreditation with organisations such as the WOAH and national testing authorities is encouraged but not mandatory (Halpin* et al., *2021; Hobbs* et al., *2020; and Section 2.5* Robustness and ruggedness *below*).

PRELIMINARY CONSIDERATIONS IN ASSAY DEVELOPMENT AND VALIDATION

All laboratories should comply with the requirements of Chapter 1.1.5 *Quality management in veterinary testing laboratories* (Newberry & Colling, 2021). 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. Comprehensive and well-designed experiments are required to develop and optimise assays with favourable analytical characteristics. The underlying principles are broadly applicable to all assay types and, when conducted with appropriate rigour, provide the foundations for high-quality diagnostic tests that are fit for their intended purpose(s) Bowden *et al.* (2021). In their review of WOAH recommended diagnostic tests, Cullinane & Garvey (2021) concluded that enzyme-linked immunosorbent assay (ELISA) and molecular assays were the most commonly used WOAH-recommended tests, which is why examples in the following sections focus on these methods (Mayo *et al.,* 2021).

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 the: (a) sample – individual or pooled, matrix composition, and host/organism interactions affecting the target analyte quantitatively or qualitatively; (b) 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) 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 used for routine testing. Integrity of laboratory-based experimental outcomes during assay development and validation is only as good as the quality of the samples used. Anticipating 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. Virtual biobanks have become relevant resources with respect to reagents and samples during test development and validation. For example, once the genomic sequence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) became available, European Virus Archive Global (EVAg) Members rapidly developed virus stocks, diagnostic tests and positive controls and, within 10 weeks, over 1500 products had been distributed worldwide for diagnostic or research purposes (Watson *et al.,* 2021). Chapter 2.2.6 *Selection and use of reference samples and panels* provides an overview about selection and use of reference samples and panels to address relevant validation parameters such as repeatability, reproducibility, DSe and DSp etc. Information on sample collection, preparation, preservation, management, and transport is available in Chapters 1.1.2 *Collection, submission and storage of diagnostic specimens* and 1.1.3 *Transport of biological materials*.

The matrix in which the targeted analyte is found (serum, faeces, tissue, etc.) may contain endogenous or exogenous inhibitors that may interfere with the performance of the assay. This is of particular concern for enzyme-dependent tests such as polymerase chain reaction (PCR) or 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

Accuracy and precision are two independent parameters which ultimately define the performance of a diagnostic test. Sensitivity, specificity and functions of these variables (e.g. likelihood ratios), change with the cut-off and represent accuracy while repeatability and reproducibility are measures of precision. A reliable test is accurate and precise. 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 (box with criteria for assay development and validation). The assay may need additional optimisation or may be found to be fit for purpose based on the results of the validation work. WOAH's Secretariat for Registration of Diagnostic Kits (SRDK) has a test validation and certification process where new tests, which are deemed to be fit for purpose(s) can be registered[2].

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 the intended purpose(s)

2 https://www.woah.org/en/what-we-offer/veterinary-products/diagnostic-kits/the-register-of-diagnostic-kits/

A. ASSAY DEVELOPMENT PATHWAY

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

Criteria for interpretation of results	
Positive predictive value	(PV+)
Negative predictive value	(PV–)
Positive likelihood ratio	(LR+)
Negative likelihood ratio	(LR–)

ISO/IEC 17025 (2017) for testing and calibration laboratories states that the laboratory shall use appropriate methods and procedures for all laboratory activities (Newberry & Colling, 2021). In other words, the assay must be 'fit for purpose'. Failure to define the purpose of testing *a priori* will likely lead to errors in both the design of the assay and in the determination of critical test parameters with the potential to invalidate the entire assay development and validation process and ultimately result in a test that does not meet the user's needs. Such errors may occur when, for example, an inappropriate reference population is selected that is heterologous to the population for which the test is being developed.

The qualitative and quantitative assessment of capacity of a positive or negative test result, e.g. predictive value and likelihood ratio to predict with high confidence the infection or exposure status of the animal or population of animals is the ultimate consideration of assay validation (Section B.4.2). This capacity is dependent on development of a carefully optimised and standardised (Section A.2.3) assay that, through accrual of validation data, provides confidence in the assay's ability to perform according to the intended purpose (Table 1).

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.

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 listed in the WOAH *Terrestrial Manual* 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).

Table 1: Test purposes and relative importance of diagnostic sensitivity (DSe), diagnostic specificity (DSp), positive predictive value (PV+), negative predictive value (PV–), likelihood ratio of a positive test result (LR+) and likelihood ratio of a negative test result (LR–)

Purpose	Examples and measures of diagnostic accuracy depending on test purpose (Reid *et al.,* 2022)
1a) Historical freedom (with or without vaccination)	In a population that is historically free of a particular disease/pathogen, the prevalence is zero (or close to zero). To be fit for purpose, the test or test algorithm aims to minimise chances of false-positive results and ideally requires **high DSp**, **high PV+ and high LR+.** This can be achieved by a single test with a high DSp or serial testing(*).
1b) Re-establishment of freedom after outbreak	During the course of a successful disease control programme, a gradual shift in prevalence from high (during the peak of the outbreak) to low (at the tail end of an outbreak) can be expected. During the early stages of a proof-of freedom testing programme, when disease prevalence remains at non-negligible levels, a fit for purpose test needs a **high DSe**, **high PV– and high LR–.** This approach minimises the chances of false-negative results and allows detection of positive individuals. This can be achieved by a single test with a high DSe or parallel testing(*). At the end of the disease control programme when the remaining infected animals have been removed from the population, disease prevalence will be very low and so the proof-of-freedom testing algorithm will likely need to be altered to increase DSp (and thus improve PV+ and LR+) similar as in 1a.
2) Certify freedom from infection or agent in individual animals or products for trade/movement purposes	For the purpose of trade and movement the probability of false-negative results needs to be minimised. Otherwise, infected animals could be traded or moved with the potential to spread infection into non-infected, healthy populations. As the test is applied on individuals, no or little information is available about the prevalence or pre-test probability of infection. To be fit for purpose the test or test algorithm aims to minimise chances of false-negative results and ideally requires **high DSe**, **high PV– and high LR–**. This can be achieved by a single test with a high DSe or parallel testing(*).
3) Contribute to the eradication of disease or elimination of infection from defined populations	This purpose follows a similar pattern as in 1b, where prevalence is expected to decrease from high to low over time in a defined population.
4) Confirm diagnosis of suspect or clinical cases (includes confirmation of positive screening test)	The goal of a confirmatory test is to minimise the chances of a false-positive result. *Confirmation of clinical cases* For the purpose of confirmation of a clinical case ideally a test with a **high DSp, high PV+** and **high LR+** is needed. Because of the clinical manifestation of the disease and expected high pathogen load DSe is not considered to be as relevant. *Confirmation of positive screening samples* Screening tests are applied on healthy populations. They usually have a high DSe to ensure infected individuals are not missed. Only if confirmed by a confirmatory test with a high DSp the animal is considered positive. In this case the confirmatory test* needs to have a **high DSp**, **high PV+** and **high LR+**. This approach follows the series testing algorithm(*).
5) Estimate prevalence of infection or exposure to facilitate risk analysis	Epidemiologists require reliable estimates for test accuracy to design sampling plans for prevalence studies, surveys, herd health status and disease control measures. Using a screening test with a high DSe followed by a confirmatory test with a high DSp is a common approach for prevalence estimations.
6) Determine immune status a) in individual animals post-vaccination b) estimate sero-prevalence post-vaccination (research and monitoring of vaccine efficacy)	For this group the aim is to have a **high DSp, PV+** and **LR+**. A false-positive result could have fatal consequences because such an animal could in fact not be vaccinated/protected. The higher the accuracy of the test the more precise will be the estimate of post-vaccine seroconversion in individuals and populations. An example is the fluorescent antibody virus neutralisation (FAVN) test to assess the immune status of dogs and cats post-vaccination against rabies virus. For international travel a result of > 0.5 IU/ml is considered to represent acceptable protection.

(*)Multiple testing

Multiple testing consists of using more than one test to determine the infection status of an animal. The most common algorithms are testing in series or in parallel. For example, if two tests are used in series a sample is considered positive only if the first and the second test are positive. Series testing increases DSp but decreases DSe and increases PV+ and LR+. Confirmatory testing follows the series testing approach because a positive result from a screening test (high DSe) needs to be confirmed by a second test with a high DSp. The confirmatory test needs to have at least the same DSe as the screening test otherwise they could generate false-negative results, which would be considered a true-negative in this algorithm. Confirmatory testing increases DSp but decreases DSe, and increases PV+ and LR+. If two tests are used in parallel a sample is considered positive if any of the tests or both tests are positive. Parallel testing increases DSe but decreases DSp and increases PV- and LR-. In addition, for multiple testing algorithms to be effective screening and confirmatory tests should not be conditionally dependent (for example measure the same analyte, such as two ELISAs using the same antigen). Positive dependence in test sensitivity reduces the sensitivity of parallel test interpretation and a positive dependence in test specificity reduces the specificity of serial interpretation (Gardner *et al.,* 2000).

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[3], which cover best practices for development and validation of assays for detection of various analytes e.g. antibody, antigen, and nucleic acid detection, Chapters 2.2.1 *Development and optimisation of antibody detection assays*, 2.2.2 *Development and optimisation of antigen detection assays*, and 2.2.3 D*evelopment and optimisation of nucleic acid detection assays*, respectively.

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.

Factors that affect the analytical characteristics of diagnostic assays are numerous and may vary according to each assay type, e.g. the main factors affecting the analytical characteristics of serological and molecular assays are described in Bowden *et al.* (2021).

For molecular assays, DSe is more dependent on the ability to obtain the target analyte in a processed sample from an animal that has the disease, than on the inherent ability of the assay to detect very low concentrations of analyte. Although an assay, such as real-time PCR, may be extremely sensitive analytically, this may not always translate to high DSe, due to the potential shortcomings of sampling (small volume, inhibitory substances, variations in the clinical spectrum of disease in an individual animal and efficiency of extraction). The inhibition of Taq polymerase due to sample characteristics may cause false-negative results in a test that otherwise has high analytical sensitivity (ASe). In addition, when evaluating two separate, real-time PCR assays for African swine fever virus (ASFV), it was found that both had comparable ASe (equivalent to 14 genomic copies of ASFV) when using a high-quality plasmid construct containing ASFV VP72 as a template. However, when evaluating various samples collected from clinically diseased pigs, the resulting DSe was determined to be 83% for one assay and 92% for the other (data not shown). The amplicon size of the assay with the lower DSe (250 base pairs or bp) was significantly larger than that of the other assay (75 bp). In general, amplification efficiency would be expected to decrease with increasing amplicon size. This characteristic becomes more pronounced when testing viral nucleic acids extracted from clinical samples, the quality of which may be affected by varying degrees of degradation. Such an outcome had not been evident when using plasmid as the template during determination of the assay's ASe. Furthermore, as an assay such as real-time PCR is so highly sensitive analytically, great care must be taken to prevent carry-over contamination with previously amplified template. Such contamination would cause a false-positive result in an assay that is otherwise also considered very highly specific (analytically) (Bowden & Wang 2021).

3 https://www.woah.org/fileadmin/Home/eng/Health_standards/tahm/2.02.00_INTRODUCTION.pdf

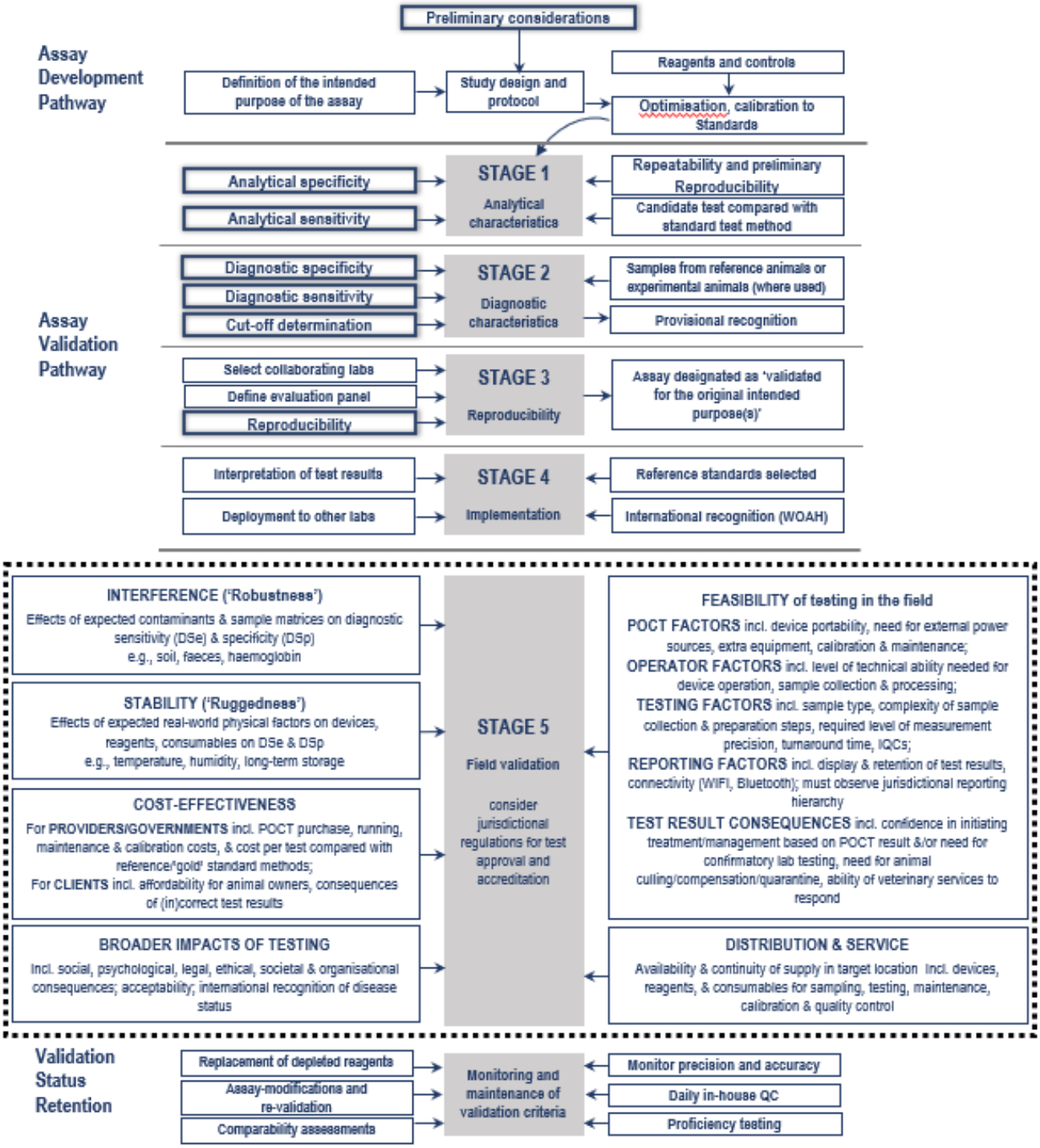

Fig. 1. Modified general assay development and validation pathways (Stages 1–4 and validation status retention) with validation criteria highlighted in bold typescript within shadowed boxes. Specific criteria for the field validation of POCTs are added in Stage 5 within the dotted box (Halpin et al., 2021); (IQCs = internal quality control sample).

2.2. 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. strong positive, weak 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 2.2.6). 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 Bowden *et al.* (2021), 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 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 2.2.8 *Comparability of assays after changes in a validated test method* for additional information on establishing comparability when reagents or processes are changed; also Reising *et al.,* 2021).

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; Watson *et al.,* 2021).

2.3. 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 is the degree of dispersion (variance, standard deviation [SD] or coefficient of variation [Cv]) within a series of measurements of the same sample tested under specified conditions. 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. The repeatability of results of operators in different laboratories is likely to be lower (higher SD, higher Cv) than that of operators working in the same laboratory. 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

usually selected (ideally, this sample will be the same one from among the three samples described in Section A.2.3 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 of pathogens, 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).

2.4. Inhibitory factors in the 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. For molecular assays, inhibitors of enzymes in the reaction mix, interferents and degradants may be present in the matrix. Selectivity is essentially the test for detection in the presence of inhibitors. Sample matrix variation is one of the most important, but may be among the least acknowledged, sources of error in analytical measurements. Assessment of ASp should use matrices relevant to the intended purpose, such as solid tissue, whole blood and swabs. Interfering substances (inhibitors) can originate from the samples, e.g. haemoglobin (Schrader *et al.,* 2012; Wilson, 1997), from the environment, or from the process of sample collection and transport (e.g. media and anticoagulants) (Druce *et al.,* 2012; Garcia *et al.,* 2002; Gibb *et al.,* 1998; Miyachi *et al.,* 1998; Yokota *et al.,* 1999;). For example, whole blood is one of the most important and common samples submitted for Hendravirus (HeV) real-time RT-PCR testing. It was found that lithium heparin, a common anticoagulant, had a significant impact on HeV real-time RT-PCR testing, producing false-negative results. In contrast, the anticoagulant ethylenediamine tetra-acetic acid (EDTA) had a comparatively low inhibitory effect. Furthermore, for the real-time RT-PCR assay for detecting bluetongue virus (BTV), diluting blood 1/10 (volume per volume or v/v) in phosphate-buffered saline resulted in improved sensitivity, evidenced by lower Ct values, in comparison to undiluted blood. A recent review of the different types of internal controls available for monitoring the inhibition of real-time PCR-based assays (Yan *et al.,* 2020) provides information on internal control strategies as a routine quality management component in veterinary molecular testing. Data collected during the validation process concerning assay performance using the sample matrices being targeted will allow for a risk-based decision as to whether an inhibition control should be included for each sample or whether the test system is unlikely to be affected by inhibition. If inhibitory substances are a significant problem, an inhibition control must be included for each test sample (Section 1.2; Bowden *et al.,* 2021).

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. 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 (Bowden *et al.,* 2021; Dejaegher & Vander Heyden, 2006). Once optimisation is complete, the robustness of the assay becomes part of the assessment of repeatability. For point-of-care tests (POCTs) (see Figure 1 Stage 5), ruggedness (is expressed as the lack of influence on test results of operational and environmental variables of the analytical method) assessments are also necessary because POCT performance should not be easily affected by operator proficiency, fluctuations in temperature, humidity, sunlight, or other environmental factors.

2.6. Calibration of the assay to standard reagents

2.6.1. International and national reference standards

Ideally, WOAH or other international reference standards, containing a known concentration or titre of analyte, are the reagents to which all assays are standardised (see WOAH Guides[4] and also chapter 2.2.6). Such standards are prepared and distributed by WOAH 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 (Watson *et al.,* 2021).

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 standards (see chapter 2.2.6). This local in-house standard therefore becomes the best available standard, and is retained in sufficient aliquoted 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, aliquoted 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, e.g. genomic copies or titre of an analyte derived from a standard curve such as a cycle threshold (Ct) value in a hydrolysis probe assay. 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 2.2.1, 2.2.2 and 2.2.3; Bowden *et al.* (2021).

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).

4 Available at: https://www.woah.org/en/what-we-offer/veterinary-products/reference-reagents/

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)". Lack of statistically robust numbers of samples from infected and non-infected animals is frequently observed with new and emerging zoonotic disease and can be a major obstacle to obtaining reliable accuracy estimates (Colling *et al.,* 2018; Stevenson *et al.,* 2021). In these circumstances tests with acceptable analytical sensitivity (ASe) and analytical specificity (ASp), repeatability and promising but preliminary DSe and DSp and reproducibility results based on a limited number of samples can be provisionally recognised by national authorities or trading partners until results from further testing confirm overall fitness for purpose. Well characterised samples have been used successfully for different purposes in interlaboratory comparison studies to assess fitness for purpose (Chapter 2.2.6; Gardner *et al.,* 2021). New platforms that can detect multiple pathogens simultaneously, e.g. multiplex technologies, high-throughput sequencing, biomarker assays and point-of-care tests represent new challenges for purpose-oriented validation studies (Bath *et al.,* 2020; Halpin *et al.,* 2021; Reid *et al*., 2021; van Borm *et al.,* 2016).

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[5], 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 2.2.6). 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. Bowden & Wang (2021) provide an example for inter-assay repeatability of three ASFV real-time PCR assays that were evaluated using a weak-positive ASFV sample. The testing, including the extraction of viral DNA, was conducted on ten different days by different operators in the same laboratory. 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 2.2.4 *Measurement uncertainty* for assessments of repeatability).

1.2. 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 2.2.1, 2.2.2 and 2.2.3 for guidance for antibody, antigen and nucleic acid assays, respectively. For example, assessing the analytical specificity of a foot and mouth disease

5 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.

(FMD) PCR could be performed by testing reference samples known to be positive for vesicular stomatitis virus, swine vesicular disease and/or malignant catarrhal fever. Analytical specificity assessments cannot determine the full range of potential cross-reacting analytes present in the population, or account for population-level sample variability; as such, determining the analytical specificity is not a surrogate for assessment of diagnostic specificity. ASp is documented during Stage 1 validation, and cross-reactions identified (Ludi *et al.,* 2021). 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); and 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 known organisms that are potentially cross-reactive. This would also define a confirmatory assay. For example, an assay to detect avian influenza virus (AIV) H5 subtypes should be assessed for cross-reaction with non-H5 AIV subtypes. Specificity testing should also include other organisms that cause similar clinical signs, to demonstrate the utility of the assay for differential detection of the target organism.

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, e.g. a group-specific bluetongue (BTV) ELISA that detects antibodies to all BTV serotypes or an NSP FMD ELISA that detects antibodies to all seven FMD serotypes .

1.3. Analytical sensitivity

The limit of detection (LOD) is a measure of the 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 2.2.5 *Statistical approaches to validation*.

1.4. Analytical accuracy of ancillary tests or procedures

Some test methods or procedures may be qualified for use as analytical tools in the diagnostic laboratory. These usually are secondary 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, molecular sequencing and MALDI-TOF-MS (matrix assisted laser desorption ionisation time of flight mass spectrometry) for bacteria (Ricchi *et al.,* 2016).

Such ancillary 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). Inthese examples, the targeted analyte is further characterised quantitatively or qualitatively by the analytical tool.

2. Stage 2 – Diagnostic performance of the assay

Diagnostic sensitivity
Percentage of known infected animals that test positive in the assay; infected animals that test negative are considered to have false-negative results.

Diagnostic specificity
Percentage of known uninfected animals that test negative in the assay; uninfected animals that test positive are considered to have false-positive results.

Animal samples used to assess DSe and DSp come from four main sources: (1) reference banks with samples of known infection status, (2) outbreak or surveillance samples where animal status is unknown but population status (infected or not infected) is known, (3) neither animal nor population status known, and (4) experimental challenge studies (see Table 3 for advantages and limitations of each method – in the Figure, groups 2 and 3 are combined for brevity). DSe (proportion of samples from known infected animals that test positive in an assay) and DSp (the proportion of samples from known uninfected animals that test negative in an assay) are the primary performance indicators established during validation of an assay (see chapters 2.2.1., 2.2.2., 2.2.3). These estimates are the basis for calculation of other parameters from which inferences are made about test results (e.g. predictive values and likelihood ratios 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 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 2.2.5.

Case definition, e.g. what constitutes an infected animal and what constitutes an uninfected animal, respectively, e.g. clinical signs, fever, results from a reference test, etc.

Reference samples: The designated number of known positive and known negative reference samples will depend on the likely values of DSe and DSp of the candidate assay and the desired confidence level for the estimates (Table 2 and Jacobson, 1998). Table 2 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 a 2% vs 5% error for a likely DSe or DSp of 90% and 95% confidence shows a considerable increase (864 *vs* 138) in the number of samples required. For the most economically important listed diseases, 99% confidence might be preferred. 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 WOAH reference standards may be lacking (see chapter 2.2.5 for further details). Therefore, it may be necessary to use fewer 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.

Table 2. 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

The following are examples of reference populations and methodologies that may aid in estimating 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 (Table 4). 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 2.2.6). After the initial detection of a novel disease, reference samples may not exist or be in limited quantities. In these situations, samples from experimental challenge studies might be the only available sample for initial validation of assays.

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 4 in Section B.2.5). Situations where a perfect reference is available for either positive or negative animals, and one where the reference is perfect for both are described for diagnostic test validation by Heuer & Stevenson (2021).

2.2. Samples from animals of unknown status

In some situations, the infection status of a population may be known. If the population is known to be non-infected, then all animals in that population are also assumed to be non-infected. In an infected population, not all animals will be infected especially if the disease is not highly contagious. Deducing the population’s status is more challenging if the infection is subclinical or covert or if animals are protected by vaccination

or prior exposure to a pathogen. When the so-called reference standard (true infection status) is imperfect, which is the rule with any diagnostic tests used to test field samples, 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 (LCA) of the joint results of two or more tests assuming neither test is perfect (e.g. Johnson *et al.,* 2019).

Latent-class models do not rely on the assumption of a perfect reference test but rather estimate the accuracy of the candidate test and a reference test using the joint test results (Branscum *et al.,* 2005; Enoe *et al.,* 2000; Georgiadis *et al.,* 2003; Hui & Walter, 1980). If a Bayesian framework is used for the LCA, prior knowledge about the DSe and DSp of the reference test and the candidate test can be incorporated into the analysis. The infection status of source populations can also be specified in a Bayesian LCA and inclusion of animal samples from known non-infected populations can enhance the ability of these models to estimate DSp and other parameters in the model (e.g. DSe and prevalence) more precisely than would occur if only data from two or more infected populations are used. Three key assumptions must normally be satisfied in a LCA: (1) when data from multiple populations are used, each population prevalence should be different; (2) the DSe and DSp of the test are constant across test populations; and (3) the tests are conditionally independent. However, if conditional dependence exists between the index and reference tests (e.g. they both measure a similar analyte), latent class models (LCM) with different dependence structure have been developed to model the conditional dependence among tests.

Because Bayesian latent class models are complex and require adherence to 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 should be based on peer-reviewed literature (see chapter 2.2.5 for details and Cheung *et al.,* 2021). Sample size calculations for a LCA require assistance from a statistician or epidemiologist with experience with the techniques.

Table 3. Source populations for test validation samples.

Reference sample banks (e.g. from Ref. Labs)	Field populations with animals of unknown infection status	Experimental challenge trials
Advantages • Infection status known • Simple statistical analysis **Limitations** • Often not available for rare or new diseases • May not be representative of animals to be tested in the future • Clinical status, results of other tests and demographic information may be missing	**Advantages** • Representative of animals to be tested in future • Latent class models (LCM) used for data analysis: infection status of population and covariates can be added • Assessment of accuracy of multiple assays is possible **Limitations** • LCM requires expertise and training, and adherence to model assumptions	**Advantages** • Infection status based on challenge group • Assessment of "diagnostic window" of an assay **Limitations** • Samples from challenge trials may not mirror samples collected in the field • Often a secondary outcome in a pathogenesis study • Route of exposure, infectious dose, and other experimental conditions can be influential • Best suited to acute infections • Ethical considerations

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 (Table 3). 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 "proof-of-concept" results and less than ideal estimates of the true DSp and DSe.

2.4. Selection of a cut-off (threshold) value for classification of test results

To obtain DSe and DSp estimates of the candidate assay, which is measured on a continuous scale, the test results must first 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; 2000; Jacobson, 1998; Zweig & Campbell, 1993; and chapter 2.2.5). 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 (reference) 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 4 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. Assume that the samples were then run with the new assay. Table 4 is a hypothetical set of results from which DSe and DSp estimates have been obtained.

Table 4. 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 2 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 2.2.5 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 2 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 2.2.5 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 2 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
Test with acceptable ASe and ASp, repeatability and promising preliminary DSe and DSp based on a small select panel of well-characterised samples containing the targeted analyte and a preliminary estimate of reproducibility. Results can be provisionally recognised until adequate samples sizes for both DSe and DSp are obtained (refer to Table 2) and a more extensive reproducibility study is done to confirm overall fitness-for-purpose.

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 2.2.6). 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

6 Provisional recognition does not imply acceptance by WOAH. 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.

can be estimated for both the reproducibility and repeatability data (see chapter 2.2.4 for further explanation of the topic and its application). Factors affecting testing reproducibility among laboratories and practical examples of proficiency testing and interlaboratory comparison testing are provided by Johnson & Cabuang (2021) and Waugh & Clark (2021). A case study with FMD for selection and use of reference panels is presented in Ludi *et al.* (2021) and the value of virtual biobanks for transparency purposes with respect to reagents and samples used during test development and validation is reported by Watson *et al.,* 2021.

For POCTs, 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.

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; Halpin *et al.,* 2021). Examples supporting each of the six purposes are provided in Table 1 *Test purposes and relative importance of diagnostic sensitivity* (*DSe*), *diagnostic specificity* (*DSp*)*, positive predictive value* (*PV+*), *negative predictive value* (*PV–*)*, likelihood ratio of a positive test result* (*LR+*) *and likelihood ratio of a negative test result* (*LR–*).

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:

$$PPV = \frac{P \; x \; DSe}{P \; x \; DSe + (1 - P) \; x \; (1 - DSp)} \quad \text{and} \quad NPV = \frac{(1 - P) \; x \; DSp}{P \; x \; (1 - DSe) + (1 - P) \; x \; DSp}$$

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 infection 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 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. Whilst predictive values can be a useful tool in diagnostic test interpretation, they are highly dependent on the prevalence of disease in the population. Predictive values calculated in populations with high prevalence, or at the peak of an outbreak where disease prevalence is high, are not applicable in populations with low prevalence or at the tail end of an outbreak where disease prevalence is markedly decreased.

The likelihood ratio (LR) indicates the diagnostic power of a given test result and can be used to assist in test interpretation. The likelihood ratio is an inherent characteristic of the test; it depends solely on the combined diagnostic sensitivity and diagnostic specificity and therefore does not vary with prevalence. Likelihood ratios are extremely powerful, as they can be used to calculate the 'post-test' probability of disease, given the observed quantitative test result, and the diagnostician's assessment of the probability of infection, prior to the test being performed (Caraguel & Colling, 2021).

The LR is calculated as the ratio of the likelihood of a given test result occurring in infected individuals to the likelihood of the same test result occurring in non-infected individuals. Conversely, if the LR is lower than one, the test result supports the absence of the infection, i.e. this test result is less likely to occur in infected animals than in non-infected animals. A LR equal to '1' means that the test result has no diagnostic power (i.e. it is as likely to occur in infected animals as it is in non-infected animals). The further the LR is away from one, towards either zero or infinity, the stronger the evidence provided by the test result. In the clinical context, test outputs with an LR > 10 or < 0.1 are considered good diagnostic evidence of the infection being either present or absent, respectively.

$$LR^{+} = \frac{DSe}{1 - DSp} \quad \text{and} \quad LR^{-} = \frac{1 - DSe}{DSp}$$

Where:

LR+ = Likelihood ratio of a positive test result

LR– = Likelihood ratio of a negative test result

DSe = Diagnostic sensitivity

DSp= Diagnostic specificity

LR ranges from zero to infinity. If the LR of a given test result is greater than one, this test result supports the presence of the infection. LRs can be applied at the test cut-off or at different ranges of results.

4.3. International recognition

Traditionally, assays have been recognised internationally by WOAH 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 WOAH procedure for validation and certification of diagnostic assays, the final step is listing of the test in the WOAH Register. Tests listed in the Register are certified as fit for a specific purpose if they have completed Validation Stages 1, 2 and 3 followed by supportive review from a panel of independent experts. 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 (Gifford *et al.,* 2021). The Register is available on the WOAH website (see footnote 2).

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 (Brown *et al.,* 2021). In the natural progression of diagnostic and/or technological improvements, recently validated assays may 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 (1.1 Repeatability). The performance can be monitored graphically by plotting measurements from assay controls in control charts[7] (Crowther *et al.,* 2006). 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 (Johnson & Cabuang, 2021) (3.1 Reproducibility). 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 (Waugh & Clark, 2021).

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

Over time, modifications of the assay will likely 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

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.

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 2.2.8 for description of experiments that are appropriate for comparability testing, chapter 2.2.6 on reference sample panels, Bowden & Wang, 2021 and Reising *et al.* 2021.

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 2.2.8 and Reising *et al*. (2021) 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.

> A **comparability study** is required when a change has been made in a test protocol of a validated test to ensure that test performance is comparable.

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. Practical experiences show that regular updates of validation dossiers with new data is difficult. Participation in proficiency testing rounds using relevant panels with the latest strains may help to assist in proving ongoing assay accuracy and precision.

5.3.1. Data management

Long-term storage of data, review of validation data and on-going verification is an important component for enhancing confidence that assays are still performing with acceptable diagnostic accuracy. To achieve that goal, Laboratory Information Management Systems (LIMS) are required that integrate validation and diagnostic data to provide regular updates of assay performance. At a minimum, data management systems should facilitate: (1) storage of validation data in central locations, (2) review and/or sharing of data when, for example, assays are deployed to external laboratories (Stage 4 of the WOAH Assay Validation Pathway (see Figure 1), (3) on-going accrual and integration of diagnostic results to update the diagnostic characteristics of an assay (Stage 3 (see Figure 1), and (4) storage of Internal Quality Control (IQC) and External Quality Control (EQC) data for use in reviewing assay sensitivity (Stage 1, (see Figure 1), especially when changes to reagents or equipment occur.

Access and utilisation of LIMS for the purposes of test validation and ongoing verification remains limited, as they require significant investment and Information Technology expertise to maintain.

6. Verification of existing validated assays

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 (Kirkland & Newberry, 2021).

A **verification study** is required when a validated test is used in a new laboratory to ensure that results from the original validation study can be verified and overall test performance is comparable.

7. New technologies

The use of high-throughput sequencing (HTS) and opportunities for its application to diagnosis are growing rapidly; the major purposes are unbiased sequencing for pathogen discovery and targeted sequencing for detection and further characterisation. If the assay is used for detecting previously unidentified microorganisms, such as during an outbreak investigation, then the primary purpose is diagnostic. If the assay is used to further characterise a previously identified pathogen or to follow the molecular epidemiology of the pathogen during an outbreak, then its general purpose can be described as an ancillary test. In the absence of a known target analyte, following a traditional validation pathway is not possible. Monitoring quality metrics such as depth of coverage, uniformity of coverage, GC bias, base-call quality scores, decline in signal intensity or read-length, mapping quality and the inclusion of internal controls are used to assess relative performance of various HTS assays (Halpin *et al.*, 2021; van Borm *et al.,* 2016).

Another complexity in diagnostic test validation is the development of multiplexed assays such as bead-based assays and multiplexed real-time RT-PCR/PCR where more than one target is identified. Test accuracy of each of those targets in different concentrations and distribution is important and challenging to validate.

8. Conclusions

Adequate validation and verification of the performance characteristics of diagnostic tests for infectious diseases is critical to ensuring that assays are applied and interpreted in a scientifically robust and defensible manner (Colling & Gardner, 2021). Since it was first adopted in 1996, the WOAH Assay Development and Validation Pathway (Figure 1) has acted as the internationally recognised standard for the validation of veterinary diagnostic tests for infectious diseases.

Whilst the WOAH standard outlines a comprehensive approach to test validation, experience has demonstrated that assay developers continue to face a number of critical challenges in complying with many stages of the pathway. In identifying these obstacles, opportunities for improvement in diagnostic test validation standards, approaches and extension activities have been identified and are summarised in Table 5 (Reid *et al.,* 2022).

Table 5. Summary of challenges and opportunities for diagnostic test validation

Challenges	Opportunities
• What is the purpose of the test?	• Clearly define the purpose(s) and relevant associated parameters, e.g. screening test requires high diagnostic sensitivity and confirmatory test requires high diagnostic specificity
• Define scope and limitations of test	• What can be expected from the test (scope) and what cannot be expected (limitations)?
• Case definition: what constitutes an infected animal and what constitutes a non-infected animal?	• For example: positive in reference test(s), characteristic lesions, experimental infection, sample taken from an individual from a historically negative population, etc. if infection status is not known Bayesian latent class model (BLCM) may be applicable
• Is reference test imperfectly accurate or likely inferior to the new test under evaluation	• Use BLCM for estimation of diagnostic sensitivity and specificity, likelihood ratios, and other relevant parameters, e.g. prevalence
• Species and specimens	• Define species and specimen for which test will be validated, e.g. domestic chicken, nasopharyngeal swabs
• Source and target population	• Is source population (where test was validated) similar to target populations (where test will be applied)?
• Design, analysis, interpretation and reporting of validation/verification studies (lack of original validation data)?	• *Terrestrial Manual* chapters 1.1.6, 2.2.1–2.2.8, WOAH certification and registration process, STARD, ParaTB, Aquatic, BLCM, ISO/IEC 17025:2017, validation templates from national and international organisations & workshops provided by WOAH Reference Laboratories and Collaborating Centres
• Lack of samples (new and emerging, rare diseases, subclinical, wildlife diseases etc.)	• Reference panels of well described samples, if available, and samples for Interlaboratory Comparison (Network collaboration, "Vetlab"), Provisional Recognition

The thoroughness with which assays should be validated might seem a daunting task for diagnostic laboratories and research teams. However, resources are available to help plan and guide validation and verification studies, including relevant chapters of the WOAH *Terrestrial Manual*, the recent special issue of the WOAH *Scientific and Technical Review* on Diagnostic Test Validation (Colling & Gardner, 2021**)** and guidance documents published by national and regional accreditation bodies. International leaders in the field of diagnostic test validation, including the WOAH and their associated Reference Laboratories and Collaborating Centres, as well as national regulatory bodies, have an important role to play in continuing to develop the standards and systems required to ensure that assay developers have the resources required, and incentives, to meet their responsibilities to perform and report well designed and transparent validation studies. End users of diagnostic tests must also be supported to take responsibility to understand and verify assay performance in their own laboratories, and clearly communicate the uncertainty associated with diagnostic test results to their stakeholders.

FURTHER READING

BATH C., SCOTT M., SHARMA P.M., GURUNG R.B., PHUENTSHOK Y., PEFANIS S., COLLING A., SINGANALLUR N., FIRESTONE S.M., UNGVANIJBAN S., RATTHANOPHART J., ALLEN J., RAWLIN G., FEGAN M. & RODONI B. (2020). Further development of a reverse-transcription loop-mediated isothermal amplification (RT-LAMP) assay for the detection of Foot-and-Mouth Disease Virus and validation in the field with use of an internal positive control. *Transbound. Emerg. Dis.,* **67**, 2494–2506. http://dx.doi.org/10.1111/tbed.13589.

BOWDEN T.R., CROWTHER J.R. & WANG J. (2021). Review of critical factors affecting analytical characteristics of serological and molecular assays. *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 53–73. doi:10.20506/rst.40.1.3208.

BOSSUYT P.M., REITSMA J.B., BRUNS D.E., GATSONIS C.A., GLASZIOU P.P., IRWIG L., LIJMER J.G. MOHER D., RENNIE D., DE VET H.C.W., KRESSEL H.Y., RIFAI N., GOLUB R.M., ALTMAN D.G., HOOFT L., KOREVAAR D.A., COHEN J.F. & STARD (Standards for Reporting Diagnostic Accuracy (STARD)) GROUP (2015). STARD. An updated list of essential items for reporting diagnostic accuracy studies. *BJM*, 351:h5527. doi: 10.1136/bmj.h5527.

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.

BROEMELING L.D. (2011a). Bayesian Methods for Medical Test Accuracy. *Diagnostics,* **1**, 1–35; https://doi.org/10.3390/diagnostics1010001.

BROEMELING L.D. (2011b). *Diagnostics*, **1**, 53–76; https://doi.org/10.3390/diagnostics1010053.

BROWN I., SLOMKA M.J., CASSAR C.A., MCELHINNEY L.M. & BROUWER A. (2021). The role of national and international veterinary laboratories. *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 159–172. doi:10.20506/rst.40.1.3215.

CARAGUEL C.G.B. & COLLING A. (2021). Diagnostic likelihood ratio – the next generation of diagnostic test accuracy measurement. *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 299–309. doi:10.20506/rst.40.1.3226.

CHEUNG A., DUFOUR S., JONES G., KOSTOULAS P., STEVENSON M.A., SINGANALLUR N.B. & FIRESTONE S.M. (2021). Bayesian latent class analysis when the reference test is imperfect. *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 271–286. doi:10.20506/rst.40.1.3224.

COLLING A. & GARDNER I.A. (eds). (2021). Diagnostic test validation science: a key element for effective detection and control of infectious animal diseases (*Rev. Sci. Tech. Off. Int. Epiz.,* **40**.The Special Issue is available at: https://doc.woah.org/dyn/portal/index.xhtml?page=alo&aloId=41245&req=21&cid=1c1f3a2e-2399-408c-946f-fb8d0c089a57.

COLLING A. & GARDNER I.A. (2021). Conclusions: Validation of tests of WOAH-listed diseases as fit-for-purpose in a world of evolving diagnostic technologies and pathogens. *Rev. Sci. Tech. Off. Int. Epiz.,* **40**, 311–317. doi: 10.20506/rst.40.1.3227

COLLING A., LUNT R., BERGFELD J., HALPIN K., MCNABB L, JUZVA S., NEWBERRY K., MORRISSY C., HLAING LOH M., CARLILE G., WAUGH C., WRIGHT L., WATSON J., EAGLES D., LOOMES C., WARNER S., DIALLO I., KIRKLAND P., BRODER C., ZUELKE K., MCCULLOUGH S. & DANIELS P. (2018). A network approach for provisional assay recognition of a Hendra virus antibody ELISA: test validation with low sample numbers from infected horses. *J. Vet. Diag. Invest.,* 30, 362–369. https://doi.org/10.1177/1040638718760102.

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**, 913–935. doi:10.20506/rst.25.3.1706.

CULLINANE A. & GARVEY M. (2021). A review of diagnostic tests recommended by the World Organisation for Animal Health Manual of Diagnostic Tests and Vaccines for Terrestrial Animals. *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 75–89. https://doi.org/10.20506/rst.40.1.3209.

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

DRUCE J., GARCIA K., TRAN T., PAPADAKIS G. & BIRCH C. (2012). Evaluation of swabs, transport media, and specimen transport conditions for optimal detection of viruses by PCR. *J. Clin. Microbiol.,* **50**, 1064–1065. doi:10.1128/JCM.06551-11.

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.

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 [https://www.springer.com/journal/12248]).

FOORD A., BOYD V., WHITE J., WILLIAMS D., COLLING A. & HEINE H. (2015). Flavivirus detection and differentiation by a microsphere array assay. *J. Virol. Methods*, **203**, 65–72.

GARCIA M.E., BLANCO J.L., CABALLERO J. & GARGALLO-VIOLA D. (2002). Anticoagulants interfere with PCR used to diagnose invasive aspergillosis. *J. Clin. Microbiol.,* **40**, 1567–1568. doi:10.1128/jcm.40.4.1567-1568.2002.

GARDNER I.A., COLLING A., CARAGUEL C.G., CROWTHER J.R., JONES G., FIRESTONE S.M. & HEUER C. (2021). Introduction: validation of tests for OIE-listed diseases as fit-for-purpose in a world of evolving diagnostic technologies. *Rev. Sci. Tec. Off. Int. Epiz*., **40**, 19–28. doi:10.20506/rst.40.1.3207.

GARDNER I.A., COLLING A. & GREINER M. (2019). Design, statistical analysis and reporting standards for test accuracy studies for infectious diseases in animals: Progress, challenges and recommendations. *Prev. Vet. Med*., **162**, 46–55. oi:10.1016/j.prevetmed.2018.10.023.

GARDNER I.A. & GREINER M. (2006). Receiver-Operating Characteristic Curves and Likelihood Ratios: Improvements over Traditional Methods for the Evaluation and Application of Veterinary Clinical Pathology Tests. *Vet. Clin. Pathol.* **35**, 8–17.

GARDNER I.A., NIELSEN S.S., WHITTINGTON R.J., COLLINS M.T., BAKKER D., HARRIS B., SREEVATSAN S., LOMBARD J.E., SWEENEY R., SMITH D.R., GAVALCHIN J. & EDA S. (2011). Consensus-based reporting standards for diagnostic test accuracy studies for paratuberculosis in ruminants. *Prev. Vet. Med*., **101**, 18–34. PMID: 21601933.

GARDNER I.A., STRYHN H., LIND P. & COLLINS M.T. (2000). Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Prev. Vet. Med*., **45**, 107–122. doi: 10.1016/s0167-5877(00)00119-7.

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.

GIBB A.P. & WONG S. (1998). Inhibition of PCR by agar from bacteriological transport media. *J. Clin. Microbiol.,* **36**, 275–276. doi:10.1128/JCM.36.1.275-276.1998.

GIFFORD G., SZABO M., HIBBARD R., MATEO D., COLLING A., GARDNER I. & ERLACHER-VINDEL E. (2021). Validation, certification and registration of certified tests and regulatory control of veterinary diagnostic test kits*. Rev. Sci. Tech. Off. Int. Epiz*., **40**, 173–188. doi:10.20506/rst.40.1.3216.

GREINER M. & GARDNER I.A. (2000). Epidemiologic Issues in the Validation of Veterinary Diagnostic Tests. *Prev. Vet. Med.,* **45**, 3–22.

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

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.

HALPIN K., TRIBOLET L., HOBBS E.C. & SINGANALLUR N.B. (2021). Perspectives and challenges in validating new diagnostic technologies. *Rev. Sci. Tech. Off. Int. Epiz*., **40**, 145–157. doi:10.20506/rst.40.1.3214.

HOBBS E., COLLING A., GURUNG R. & ALLEN J. (2020). The potential of diagnostic point-of-care tests (POCTs) for infectious and zoonotic animal diseases in developing countries: technical, regulatory and sociocultural considerations. *Transbound Emerg Dis.,* **68**, 1835–1849.

HEUER C. & STEVENSON M.A. (2021). Diagnostic test validation studies when there is a perfect reference standard. *Rev. Sci. Tech. Off. Int. Epiz*., **40**, 261–270. doi:10.20506/rst.40.1.3223.

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

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (2019). ISO/TS 22583:2019(en), Guidance for supervisors and operators of point-of-care testing (POCT) devices. https://www.iso.org/obp/ui/#iso:std:iso:ts:22583:ed-1:v1:en

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

JIA B., COLLING A., STALLKNECHT D.E., BLEHERT D., BINGHAM J., CROSSLEY B., EAGLES D. & GARDNER I.A. (2020). Validation of laboratory tests for infectious diseases in wild mammals: review and recommendations. *J. Vet. Diagn. Invest.,* **32**, 776–792. doi:10.1177/1040638720920346.

JOHNSON P. & CABUANG L. (2021). Proficiency testing and ring trials. *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 189–203. https://doi.org/10.20506/rst.40.1.3217

JOHNSON W.O., JONES G. & GARDNER I. (2019). Gold standards are out and Bayes is in: implementing the cure for imperfect reference tests in diagnostic accuracy studies. *Prev. Vet. Med*., **167**, 113–127. doi:10.1016/j.prevetmed.2019.01.010.

KIRKLAND P.D. & NEWBERRY K.M. (2021). Your assay has changed – is it still 'fit for purpose'? What evaluation is required? *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 205–215. https://doi.org/10.20506/rst.40.1.3218.

KOSTOULAS P., GARDNER I.A., ELSCHNER M.C., DENWOOD M.J., MELETIS E. & NIELSEN S.S. (2021). Examples of proper reporting for evaluation (Stage 2 validation) of diagnostic tests for diseases listed by the World Organisation for Animal Health, *Sci. Tech. Off. Int. Epiz.*, **40**, 287–298. doi:10.20506/rst.40.1.3225.

KOSTOULAS P., NIELSEN S.S., BRANSCUM A.J., JOHNSON W.O., DENDUKURI N., DHAND N.K., TOFT N. & GARDNER I.A. (2017). STARD-BLCM: Standards for the Reporting of Diagnostic accuracy studies that use Bayesian Latent Class Models. *Prev. Vet. Med*., **138**, 37–47. PMID: 28237234.

LUDI A.B., MIOULET V., BAKKALI KASSIMI L., LEFEBVRE D.J., DE CLERCQ K., CHITSUNGO E., NWANKPA N., VOSLOO W., PATON D.J. & KING D.P. (2021). Selection and use of reference panels: a case study highlighting current gaps in the materials available for foot and mouth disease. *Rev. Sci. Tech. Off. Int. Epiz*., **40**, 239–251. doi:10.20506/rst.40.1.3221

MAYO C.E., WEYER C.T., CARPENTER M.J., REED K.J., RODGERS C.P., LOVETT K.M., GUTHRIE A.J., MULLENS B.A., BARKER C.M., REISEN W.K. & MACLACHLAN N.J. (2021). Diagnostic applications of molecular and serological assays for bluetongue and African horse sickness. *Rev. Sci. Tech. Off. Int. Epiz*., **40**, 91–104. doi:10.20506/rst.40.1.3210.

MICHEL A.L., VAN HEERDEN H., PRASSE D., RUTTEN V., DAHOUK S. AL & CROSSLEY B.M. (2021). Pathogen detection and disease diagnosis in wildlife: challenges and opportunities. *Rev. Sci. Tech. Off. Int. Epiz*., **40**, 105–118. doi:10.20506/rst.40.1.3211.

MIYACHI H., MASUKAWA A., OHSHIMA T., FUSEGAWA H., HIROSE T., IMPRAIM C. & ANDO Y. (1998). Monitoring of inhibitors of enzymatic amplification in polymerase chain reaction and evaluation of efficacy of RNA extraction for the detection of hepatitis C virus using the internal control. *Clin. Chem. Lab. Med.,* **36**, 571–575. doi:10.1515/CCLM.1998.098.

NEWBERRY K.M. & COLLING A. (2021). Quality standards and guidelines for test validation for infectious diseases in veterinary laboratories? *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 227–237. doi:10.20506/rst.40.1.3220.

PÉREZ L.J., LANKA S., DESHAMBO V.J. , FREDRICKSON R.L. & MADDOX C.V. (2020). A validated multiplex Real Time PCR assay for the diagnosis of infectious Leptospira spp.: a novel assay for the detection and differentiation from both pathogenic groups I and II. *Front. Microbiol.,* **11**, 457. https://doi.org/10.3389/fmicb.2020.00457.

REID T., SINGANALLUR N. B., NEWBERRY K., WAUGH C., BOWDEN T. & COLLING A. (2021). Validation of diagnostic tests for infectious diseases: challenges and opportunities. International Symposium on Sustainable Animal Production and Health Current Status and Way Forward. 28 June–2 July 2021, Joint FAO/IAEA Centre. Accepted for publication in symposium proceedings 7 April 2022.

REISING M.M., TONG C., HARRIS B., TOOHEY-KURTH K.L., CROSSLEY B., MULROONEY D., TALLMADGE R.L., SCHUMANN K.R., LOCK, A.B. & LOIACONO C.M. (2021). A review of guidelines for evaluating a minor modification to a validated assay. *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 217–226. doi:10.20506/rst.40.1.3219.

RICCHI M., MAZARELLI A., PISCINI A., DI CARO, A, CANNAS A., LEO S., RUSSO S. 1 ARRIGONI N. (2016). Exploring MALDI-TOF MS approach for a rapid identification of *Mycobacterium avium* ssp. paratuberculosis field isolates. *J. Appl. Microbiol.*, **122**, 568-577.

SAAH A.J. & HOOVER D.R. (1997). 'Sensitivity' and 'Specificity' Reconsidered: The Meaning of These Terms in Analytical and Diagnostic Settings. *Ann. Internal Med.*, **126**, 91–94.

SCHRADER C., SCHIELKE A., ELLERBROEK L. & JOHNE R. (2012). PCR inhibitors – occurrence, properties and removal. *J. Appl. Microbiol*., **113**, 1014–1026. doi:10.1111/j.1365-2672.2012.05384.x.

STEVENSON M., HALPIN K. & HEUER C. (2021). Detection of emerging infectious zoonotic diseases. *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 119–130. doi:10.20506/rst.40.1.3212.

TRIBOLET, L., KERR, E. COWLED, C., BEAN, A.G., STEWART, C.R., DEARNLEY, M. AND FARR, R. (2020. MicroRNA Biomarkers for infectious diseases: from basic research to bionsensing. *Front. Microbiol.,* **11**, 1197. doi: 10.3389/fmicb.2020.01197.

VAN BORM S., WANG J., GRANBERG F. & COLLING A. (2016). Next-generation sequencing workflows in veterinary infection biology: towards validation and quality assurance. *Rev. Sci. Tech. Off. Int. Epiz.*, **35**, 67–81. doi:10.20506/rst.35.1.2418.

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**, 1381–1386.

WATSON J.W., CLARK G.A. & WILLIAMS D.T. (2021). The value of virtual biobanks for transparency purposes with respect to reagents and samples used during test development and validation. *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 253–259. doi:10.20506/rst.40.1.3222.

WAUGH C. & CLARK G. (2021). Factors affecting test reproducibility among laboratories. *Rev. Sci. Tech. Off. Int. Epiz.*, **40**, 131–143. doi:10.20506/rst.40.1.3213.

WILSON I.G. (1997). Inhibition and facilitation of nucleic acid amplification. *Appl. Environ. Microbiol.*, **63**, 3741–3751. doi:10.1128/AEM.63.10.3741-3751.1997.

YAN L., TOOHEY-KURTH K.L., CROSSLEY B.M., BAI J., GLASER A.L., TALLMADGE R.L. & GOODMAN L.B. (2020). Inhibition monitoring in veterinary molecular testing. *J. Vet. Diagn. Invest.,* **32**, 758–766. doi:10.1177/1040638719889315

YOKOTA M., TATSUMI N., NATHALANG O., YAMADA T. & TSUDA I. (1999). Effects of heparin on polymerase chain reaction for blood white cells. *J. Clin. Lab. Anal.,* **13**, 133–140. doi:10.1002/(sici)1098-2825(1999)13:3<133::aid-jcla8>3.0.co;2-0.

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

https://www.iaea.org/services/networks/vetlab

https://www.iaea.org/services/zodiac

https://www.woah.org/en/what-we-offer/veterinary-products/diagnostic-kits/the-register-of-diagnostic-kits/

https://www.awe.gov.au/agriculture-land/animal/health/laboratories/tests/test-development

*
* *

NB: There is a WOAH Collaborating Centre for
Diagnostic Test Validation Science in the Asia-Pacific Region (please consult the WOAH Web site:
https://www.woah.org/en/what-we-offer/expertise-network/collaborating-centres/#ui-id-3).
Please contact the WOAH Collaborating Centre for any further information on validation.

NB: FIRST ADOPTED IN 1996 AS PRINCIPLES OF VALIDATION OF DIAGNOSTIC ASSAYS FOR INFECTIOUS DISEASES.
MOST RECENT UPDATES ADOPTED IN 2023.