CHAPTER 2.2.6.

SELECTION AND USE OF REFERENCE SAMPLES AND PANELS

INTRODUCTION

The WOAH Validation Recommendations provide detailed information and examples in support of Chapter 1.1.6 Validation of diagnostic assays for infectious diseases of terrestrial animals*. Reference samples and panels are essential from the initial proof of concept in the development laboratory through to the maintenance and monitoring of assay performance in the diagnostic laboratory and all of the stages in between. The critical importance of reference samples and panels cannot be over-emphasised. The wrong choice of reference materials can lead to bias and flawed conclusions right from development through to validation and use. Therefore, care must be exercised in selecting reference samples and designing panels.*

***Fig. 1. Reference samples and panels grouped based on similar characteristics and composition. The topics and alphanumeric subheadings (e.g. Proof of concept, A.2.1) refer to the relevant section in Chapter 1.1.6* Validation of diagnostic assays for infectious diseases of terrestrial animals.**

Group A	Group B	Group D
Proof of concept, A.2.1.	ASp, B.1.2.	Standard method comparison, B.5.2.1.
Operating range, A.2.3.	Analytical accuracy, ancillary tests B.1.4.	Provisional recognition, B.2.6.
ASe, B.1.3.	Group C	Biological modifications, B.5.2.2.
Optimisation, A.2.2.	Repeatability B.1.1.	Group E
Preliminary repeatability, A.2.8.	Preliminary reproducibility, B.2.6.	DSp and DSe Gold standard, B.2.1.
Calibration and process control, A.2.6.	Reproducibility, B.3.	Group F
Technical modifications, B.5.2.1.	Proficiency testing, B.5.1.	DSp and DSe no gold standard, B.2.2.
Reagent replacement, B.5.2.3.		

ASp = Analytical specificity; ASe = Analytical sensitivity; DSp = diagnostic specificity; DSe = diagnostic sensitivity

Reference samples or panels are mentioned throughout chapter 1.1.6. Reference materials are "substances whose properties are sufficiently homogenous and well established to be used for the calibration of an apparatus, the assessment of a measurement method, or for assigning values to materials" [1]*. In the context of test method validation, reference materials or samples contain the analyte of interest in varying concentrations or reactivities and are used in developing and evaluating the candidate assay's analytical and diagnostic performance. Analyte means the specific component of a test sample that is detected or measured by the test method, e.g. antibody, antigen or nucleic acid. Reference samples may be sera, fluids, tissues, excreta, feed or environmental samples that contain the analyte of interest and are usually harvested from infected animals and their environments. However, in some cases, they may be prepared in the laboratory from an original starting material (e.g. a dilution of a high positive serum in negative serum) or perhaps created by spiking the chosen matrix with a derived analyte (e.g. a bacterial or viral culture, a recombinant/expressed protein, or a genomic construct). Whether natural or prepared, they are used in experiments throughout the development process, carry over into the validation pathway and can be used to monitor performance throughout the lifespan of the assay.*

1 https://www.techlab.fr/Commun/UK_Def_MRC.asp

Wherever possible, large quantities of reference samples should be collected or prepared and preserved for long-term use. Switching reference samples during the validation process introduces an intractable variable that can severely undermine interpretation of experimental data and the integrity of the development and validation process. For assays that may target multiple species, the samples should be representative of the primary species of interest. It is critical that these samples reflect both the target analyte and the matrix in which it is found in the population for which the assay is intended. The reference materials should appropriately represent the range of analyte concentration to be detected by the assay.

Whether reference samples are selected from natural sources or prepared in the laboratory, all selection criteria and preparation procedures, as well as testing requirements, need to be fully described and put into document control. Not only is this good quality management practice, but it will provide both an enhanced level of continuity and confidence throughout the lifespan of the assay. Summaries of the data to be collected and documented for reference material can be found in Figure 2. For more detail on best practice and quality standards for the documentation of provenance of reference material refer to Watson et al. *(2021).*

Fig. 2. Documentation of reference material should be thorough to ensure i) transparency of intended purpose during assay development; ii) the correct sample types are used in all stages of assay development and validation; iii) accurate replacement of depleted reagents; and iv) appropriate choice of reference material during assay modification and re-validation. Minimum descriptive metadata are listed for pathogen, animal host, tissue type and phase of infection.

<table>
<tr><th>Pathogen data</th><th>Animal host and sample type data</th><th>Phase of infection data</th></tr>
<tr><td rowspan="2">Strain/isolate
Serotype
Genotype
Lineage
Tests used for characterisation</td><td>Natural infection
Experimental infection and protocol used
Species
Breed
Age
Sex
Reproductive status
Vaccination history
Herd history</td><td rowspan="2">Clinical signs
infection status and disease outcome
Antibody profiles
Pathogen load and shedding
Tests used to determine status of disease/infection (case definition)
Time post-experimental infection/exposure</td></tr>
<tr><td>Tissue type/s (matrix) used
For spiked samples – detail source of analyte and diluent (matrix) used
Details relating to pooling of samples</td></tr>
</table>

A. GROUP A

The question of pooling of samples to create a reference sample is often asked. If reference material is harvested from a single animal, it is important to ascertain whether or not it is representative of a typical course and stage of infection within the context of the population to be tested. If not, this could lead to bias and flawed conclusions related to validation. Pooling is a good alternative but it is imperative to pool from animals that are in a similar phase of infection. This is particularly important for antibody detection systems. Pooling also addresses the issue of the larger quantities of reference material to be stored for long term use, especially when dealing with smaller host species. Before pooling any samples, it is preferable that they be independently tested to demonstrate that they are similar with respect to analyte concentration and/or reactivity. There should be an assessment following pooling to ensure that unforeseen interference is not introduced by the pooling of multiple samples, for example differing blood types or antibody composition within the independent samples could cross-react within the pool, thus causing the pooled sample to behave differently in the specified assay than the individual samples when tested independently.

It is often difficult to obtain individual samples that truly represent analyte concentrations or reactivities across the spectrum of the expected range. Given the dynamics of many infections or responses to pathogens, intermediate ranges are often very transient. In the case of antibody responses, early infection phases in individual animals often result in highly variable and heterogeneous populations of antibody isotypes and avidities. In general, these do not

make good reference samples for assessing the analytical characteristics of an assay. They are nonetheless important for different types of reference panels as will be discussed later. For most applications in Group A, it is acceptable to use prepared samples that are spiked with known concentrations of analyte or a dilution series of a high positive in negative matrix to create a range of concentrations.

Whether natural or prepared, reference samples should represent the anticipated range of analyte concentrations, from weak to strong positive, which would be expected during a typical course of infection. A negative reference sample should be included as a background monitor. If a negative (matrix) is used as diluent for preparation of a positive reference sample (e.g. a negative serum used to dilute a high positive serum or tissue spiked with a construct), that negative should definitely be included as the negative reference sample.

Above all else, natural or prepared, reference materials must be unequivocal with respect to their status as representing either a true positive or a true negative sample. This may require that the status be confirmed using another test or battery of tests. For example, many antibody reference sera are characterised using multiple serological tests. This provides not only confidence but additional documented characteristics that may be required when attempting to replace or duplicate this reference material in the future.

Recommendations regarding stability and storage of reference materials are available: https://www.woah.org/en/what-we-offer/veterinary-products/#ui-id-4

1. Proof of concept (Chapter 1.1.6, Section A.2.1)

Chapter 1.1.6 states that test methods and related procedures must be appropriate for specific diagnostic applications in order for the test results to be of relevance. In other words, the assay must be 'fit for purpose'. Many assays are developed with good intentions but without a specific application in mind. At the very outset, it is critical that the diagnostic purpose(s) should be defined with respect to the population(s) to be tested. The most common purposes are listed in broad terms in Section A of chapter 1.1.6. As such, they are inclusive of more narrow and specific applications. However, these specific purpose(s) need to be clearly defined from the outset and are critically important in the context of a fully validated assay. As will be seen in the following descriptions, clearly defining the application will have impact on both the selection of reference samples and panels and the design of analytical and diagnostic evaluations.

2. Operating range (Chapter 1.1.6, Section A.2.3) and analytical sensitivity (Chapter 1.1.6, Section B.1.3)

2.1. Operating range and analytical sensitivity

The operating range of the assay defines the lower and upper analyte detection limits and the interval over which the method provides suitable accuracy and precision. The operating range is established by serial dilution, to extinction, of replicates of a strong positive reference sample, either natural or prepared. Dilutions of the strong positive are made in negative matrix representative of the typical sample type taken from animals in the population targeted by the assay. This includes antibody assays where replicates of a strong positive reference serum should be diluted in a negative reference serum to create the dilution series. Analytical sensitivity (ASe) is measured by replicates of the lower limit of detection (LOD) of an analyte in an assay. The same strong positive reference sample may be used to determine both the operating range and the analytical LOD.

2.2. Comparative approaches to analytical sensitivity

If the intended purpose is to detect low levels of analyte or subclinical infections, it may be difficult to obtain the appropriate reference materials from early stages of the infection process. In some cases, it may be useful to determine a comparative ASe by running a panel of samples on the candidate assay and on another independent assay. Ideally this panel of samples would be serially collected from either naturally or experimentally infected animals and should represent infected animals early after infection, through to the development of clinical or fulminating disease, if possible. This would provide a relative comparison of ASe between the assays and a temporal comparison of the earliest point of detection relative to the pathogenesis of the disease.

An experiment like the one described above, provides a unique opportunity to collect reference samples representing a natural range of concentrations that would be useful for other validation purposes. Care must be taken to avoid use of such samples when inappropriate (consult Group D below). Wherever possible serial samples should be collected from a statistically sound number of animals throughout the course of infection. In cases where sampling is lethal (e.g. requiring the harvest of internal organ tissues), the number of animals required depends on need and fitness of the experimental approach. In all cases approval from an ethics committee is required. For smaller host species, the number may need to be increased in order to collect sufficient reference material. Given that experiments like this require a high commitment of resources, it would be wise to maximise the collection of not only the currently targeted reference samples but additional materials (e.g. multiple tissues, fluids, etc.) that may be useful as reference materials in the future.

3. Optimisation (Chapter 1.1.6, Section A.2.2) and preliminary repeatability (Chapter 1.1.6, Section A.2.8)

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. At least three reference samples representing negative, weak and strong positive may be chosen from either natural or prepared reference samples. Optimisation experiments are rather exhaustive especially when assays with multiple preparatory and testing steps are involved. It is very important that a sufficient quantity of each reference sample be available to complete all optimisation experiments. Changing reference samples during the course of optimisation is not recommended as this will result in the addition of an uncontrolled variable and a disruption in the continuity of optimisation evidence.

Assessment of repeatability should begin during assay development and optimisation stages and is further verified during Stage 1 of assay validation (Section B.1.1 of chapter 1.1.6). The same reference samples should be used throughout to provide continuity of evidence.

4. Calibration and process controls (Chapter 1.1.6, Section A.2.6)

4.1. International, national or in-house analyte reference standards

International reference standards are highly characterised, contain defined concentrations of analyte, and are usually prepared and held by international reference laboratories. They are the reagents to which all assays and/or other reference materials should be standardised. National reference standards are calibrated by comparison with an international standard reagent whenever possible. In the absence of an international standard, a national reference standard may be selected or prepared and it then becomes the standard of comparison for the candidate assay. In the absence of both of the above, an in-house standard should be selected or prepared by the development laboratory within the responsible organisation. In all cases, thorough documentation of reference material should be observed as summarised in Figure 2. All of the standard reagents, whether natural or prepared, must be 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). These reference standards should also be both stable and innocuous.

Reference standards, especially antibody, are usually provided in one of two formats. They may be provided as a single positive reagent of given titre with the expectation that the candidate assay will be standardised to give an equivalent titre. This is a straight forward analytical approach but many of these ‘single’ standards have been prepared from highly positive samples as a pre-dilution in a negative matrix in order to maximise the number of aliquots available. The drawback here is that there is no accounting for any potential matrix effect in the candidate assay as there is no matrix control provided. The other approach is to provide a negative and a weak and strong positive set of reference standards that are of known concentrations or reactivities and are within the operating range of the standard method that was used to prepare them. The negative provided in the set must be the same as the negative diluent used to prepare the weak and strong positive reference standard, if the positive standards were diluted. This compensates for any potentially hidden matrix effect. In addition, this set of three acts as a template for the selection and/or preparation of process controls (discussed below).

Classically, the above standards usually have been polyclonal antibody standards and to a lesser extent, conventional antigen standards used for calibration of serological assays. However, today, reference standards could also be monoclonal antibodies or recombinant/expressed proteins or genomic constructs, if they are to be used to calibrate assays to a single performance standard.

4.2. Working standards or process controls

Working standard reagent(s), commonly known as quality or process controls, are calibrated to international, national, or in-house standard reagents. They are selected or prepared in the local matrix which is found in the population for which the assay is intended. Ideally, negative and weak and strong positive working standards should be selected or prepared. Concentrations and/or reactivities should be within the normal operating range of the assay. Large quantities should be prepared, aliquoted and stored for routine use in each diagnostic run of the assay. The intent is that these controls should mimic, as closely as possible, field samples and should be handled and tested like routine samples. They are used to establish upper and lower control limits of assay performance and to monitor random and/or systematic variability using various control charting methods. Their daily performance will determine whether or not an assay is in control and if individual runs may be accepted. As such, these working reference samples are critically important from a quality management standpoint.

5. Technical modifications (Chapter 1.1.6, Section B.5.2.1)

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 may be done to determine if these minor modifications to the assay protocol will affect the test results. Consult See chapter 2.2.8 for description of experiments and statistical approaches that are appropriate for comparability testing (Bowden & Wang, 2021; Reising *et al.,* 2021).

In general, these approaches require the use of three reference samples, a negative, a weak and strong positive to represent the entire operating range of both assays. Samples may be either natural or prepared. The important point to re-iterate here is that the same reference samples that were used in the developmental stages of the assay may be used to assess modifications after the method has been put into routine diagnostic use. This provides a higher level of confidence assessing potential impacts because the performance characteristics of these reference samples have been well characterised. At the very least, if new reference samples are to be used, they should be selected or prepared using the same criteria or preparation procedures established for previous materials as this enhances the continuity of evidence.

6. Reagent replacement (Chapter 1.1.6, Section B.5.2.3)

When a reagent such as a process control sample is nearing depletion, it is essential to prepare and repeatedly test a replacement before such a control is depleted. The prospective replacement 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 control reagent at a time to avoid the compound problem of evaluating more than one variable.

Replacement reference reagent should be selected or prepared using the same criteria or preparation procedures established for previous materials as this enhances the continuity of evidence and confidence in the assay and underlines the importance of documentation of reference material data (Figure 2).

B. GROUP B

1. Analytical specificity (Chapter 1.1.6, Section B.1.2)

Analytical specificity (ASp) is the degree to which the assay distinguishes between the target analyte and other components that may be detected in the assay. The choice of reference samples that are required to assess ASp is highly dependent on the intended purpose or application defined at the development stage of the assay. Assessment of ASp is a crucial element in proof of concept and verification of fitness for purpose and may be broken down into three elements: selectivity, exclusivity and inclusivity.

Selectivity: an important element is the extent to which a method can accurately detect and or quantify the targeted analyte in the milieu of nucleic acids, proteins and/or antibodies in the test matrix. This is sometimes termed 'selectivity'. An example is the use of reference samples for tests that are designed to differentiate infected from vaccinated animals (DIVA tests).

Reference samples need to be selected and tested from i) non-infected/non-vaccinated, ii) non-infected/vaccinated, iii) infected/non-vaccinated, and iv) infected/vaccinated animals. These samples may be collected under field conditions but it is important that an accurate history be collected, ideally with respect to the animals, but at least to the herds involved, including vaccination practices and disease occurrences (Figure 2). Alternatively, it may be necessary to produce this material in experiments like those described in Section A.2.2 of this chapter, including a combination of experimentally vaccinated and challenged animals. Application of the 3 Rs (replacement, reduction and refinement) aims to avoid or minimise the number of animals used in experiments. For enzyme-linked immunosorbent assays (ELISAs), it is important to avoid use of the vaccine as capture antigen in the assay (e.g. indirect ELISA), because carrier proteins in the vaccine may stimulate non-specific antibody responses in vaccinated animals that may be detected in ELISA, leading to false positives in the assay. Depending on the DIVA test, a single experiment could be designed to assess aspects of both ASe and ASp.

Exclusivity is the capacity of the assay to detect an analyte or genomic sequence that is unique to a targeted organism, and excludes all other other known organisms that are potentially cross-reactive. This is especially true in serological assays where there are many examples of antigens expressed by other organisms that are capable of eliciting cross-reacting antibody. An attempt should be made to obtain reference samples from documented cases of infections or organisms that may be cross-reactive. Depending on the type of assay, these reference materials may represent the organism itself, host-derived samples, or genomic sequences. A profile for the exclusivity of the assay should be established, and expanded on a continual basis as potentially cross-reactive organisms arise.

Inclusivity relates to the capacity of an assay to detect one or several strains or serovars of a species, several species of a genus, or a similar grouping of closely related viruses, bacteria or antibodies. This defines the scope of detection and thus the fitness for purpose. Reference samples are required to define the scope of the assay. If for example an assay is developed as a screening test to detect all known genotypes or serotypes of a virus, then reference samples from each representative type should be tested. As new lineages or serotype variants arise, they too should be tested as part of the test profile, which should be updated on an ongoing basis.

2. Analytical accuracy of ancillary tests (Chapter 1.1.6, Section B.1.4)

Some test methods or procedures are solely analytical tools used to further characterise an analyte that has been detected in a primary assay. Examples are the virus neutralisation test used to type an isolated virus or characterise an antibody response and subtyping of haemagglutinin genes by polymerase chain reaction of avian influenza virus. Such ancillary tests must be validated for analytical performance characteristics and differ to routine diagnostic tests because they do not require validation for diagnostic performance characteristics. The analytical accuracy of these tests is often dependant on the use of reference material. These reagents, whether they are antibody for typing strains of organisms or reference strains of the organism, etc., should be thoroughly documented, as required for any other reference material (Figure 2), with respect to their source, identity and performance characteristics.

C. GROUP C

Reference samples in Group C may be used for a number of purposes. In the initial development stages, they may be used in the assessment of assay repeatability and both preliminary reproducibility in Stage 1 and the more in depth assessment of reproducibility in Stage 3 of the Validation Pathway. However, these samples have a number of other potential uses once the assay is transferred to the diagnostic laboratory. They may be used as panels for training and qualifying of analysts, and for assessing laboratory proficiency in external ring testing programmes. Ideally, 20 or more individual samples should be prepared in large volumes. About a quarter (25%) should be negative samples and the remainder (75%) should represent a collection of positives spanning the operating range of the assay. They should be aliquoted into individual tubes in sufficient volumes for single use only and stored for long term use (Chapter 1.1.2 *Collection, submission and storage of diagnostic specimens*). The number of aliquots of each that will be required will depend on how many laboratories will be using the assay on a routine diagnostic basis and how often proficiency testing is anticipated. Ideally, they should be prepared in an inexhaustible quantity, but this is seldom feasible. At a minimum, several hundred or more aliquots of each should be prepared at a time if the assay is intended for use in multiple laboratories. This allows assessment of laboratory proficiency by testing

the same sample over many testing intervals – a useful means of detecting systematic error (bias) that may creep into long term use of an assay.

These samples may be natural or prepared from either single or pooled starting material. The intent is that they should mimic as closely as possible a true test sample. Because mass storage is always a problem, it may be necessary to store these materials in bulk and prepare working aliquots from time to time. However, if storage space is available, it is preferable to prepare and store large numbers of aliquots at one time because bulk quantities of analyte, undergoing freeze–thaw cycles to prepare a few aliquots at a time, may be subject to degradation. Because this type of reference material is consumed at a fairly high rate, they will need to be replaced or replenished on a continual basis. As potential replacement material is identified during routine testing or during outbreaks, it is advisable to work with field counterparts to obtain bulk reference material and store it for future use. Alternatively, it may be necessary to produce this material in experiments like those described in Section A.2.2 of this chapter. Similar to the comparative approach described above with respect to ASe, at least five animals in each group should be considered. For smaller host species, this number may need to be increased in order to collect sufficient reference material.

1. Repeatability (Chapter 1.1.6, Section B.1.1) and provisional assay recognition (Chapter 1.1.6, Section B.2.6)

Repeatability is the level of agreement between results of replicates of a sample both within and between runs of the same test method in a given laboratory. Repeatability is estimated by evaluating variation in results of replicates from a minimum of three (preferably five) samples representing analyte activity within the operating range of the assay. Consult Chapter 2.2.4 *Measurement uncertainty* for statistical approaches for measures of uncertainty for assessments of repeatability.

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. However, preliminary reproducibility estimates of the candidate assay should be determined during developmental stages. A small panel of three (but preferably five) representing negative, weak and strong positives, like those described above, would be adequate. This type of panel could also be used for a limited evaluation of reproducibility to enhance provisional acceptance status for the assay. The test method is usually assessed in two or more laboratories with a high level of experience and proficiency in assays similar to the candidate assay. The panel of 'blind' samples is evaluated using the candidate assay in each of these laboratories, using the same protocol, same reagents and comparable equipment. This is a scaled-down version of Stage 3 of assay validation.

2. Reproducibility (Chapter 1.1.6, Section B.3)

Reproducibility is an important measure of the precision of an assay when used in a cross-section of laboratories located in distinct or different regions or countries using the identical assay (protocol, reagents and controls). As the number of laboratories increases, so do the number of variables encountered with respect to laboratory environments, equipment differences and technical expertise. An overview of the factors affecting testing reproducibility is provided in Waugh & Clark (2021). Reproducibility studies are a measure of an assay's capacity to remain unaffected by substantial changes or substitutions in test conditions anticipated in multi-laboratory use (e.g. shipping conditions, technology transfer, reagents batches, equipment, testing platforms and/or environments). At least three laboratories should test the same panel of 'blind' samples containing a minimum of 20 samples, representing negative and a range of positive samples. If selected negative and/or positive samples are duplicated in the panel then it may be possible to assess both assay reproducibility and within-laboratory repeatability.

3. Proficiency testing (Chapter 1.1.6, Section B.5.1)

A validated assay in routine use in multiple laboratories needs to be continually monitored to ensure uniform performance and provide overall confidence in test results. This is assessed through external quality assurance programmes. Proficiency testing is one measure of laboratory competence derived by means of an inter-laboratory comparison; implied is that participating laboratories are using the same (or similar) test methods, reagents and controls. Results are usually expressed qualitatively, i.e. either negative or positive, to determine pass/fail criteria. However, where semi-quantitative results are provided, additional analysis may assess non-random error among the participating laboratories. Refer to Johnson & Cabuang (2021) for an overview of proficiency testing and ring trials.

Proficiency testing programmes are varied depending on the type of assay in use. For single dilution type assays, panel sizes vary but a minimum of five samples, representing negative weak and strong positives, would be adequate.

D. GROUP D

Reference samples in Group D differ from the previous Groups in that each sample in the panel should be from a different individual animal. As indicated in Chapter 2.2.8 *Comparability of assays after changes in a validated test method*, experimental challenge studies often include repeated sampling of individual animals to determine the progression of disease, but this is a different objective to comparing performance characteristics that would be associated with diagnostic sensitivity (DSe) and diagnostic specificity (DSp) of a test method. Serially drawn samples, taken on different days from the same animal, cannot be used as representative of individual animals in populations targeted by the assay, because such samples violate the rule of independence of samples required for such studies.

Care must be taken in choosing the reference samples and the standard (independent) method used in this type of comparison to ensure that the analytes being detected (if different) demonstrate the same type of pathogenic profile in terms of time of appearance after exposure to the infectious agent, and relative abundance in the test samples chosen.

1. Standard method comparison and provisional recognition (Chapter 1.1.6, Sections B.5.2.1 and B.2.6)

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 exotic diseases). However, a small but select panel of highly characterised test samples representing the range of analyte concentration should be run in parallel in the candidate assay method and a WOAH standard method, as published in the WOAH *Manual*. Biobanks may be a useful resource in this context, providing well-characterised samples supported with metadata to enhance transparency and provenance of samples used in method comparisons (Watson *et al.*, 2021). If the methods are deemed to be comparable (Chapter 2.2.8), and depending on the intended application of the assay, the choice may be made that further diagnostic validation is not required. For example, if the intended application is for screening of imported animals or animal products for exotic pathogens or confirmation of clinical signs, full validation beyond a test method comparison may not be feasible or warranted.

Experience has shown that the greatest obstacle to continuing through Stage 2 of the Validation Pathway is the number of defined samples required to estimate diagnostic performance parameters with a high degree of certainty (chapter 1.1.6, Section B.2). In some cases, provisional recognition by international, national or local authorities may be granted for an assay that has not been completely evaluated past analytical stages. The different rationales for provisional acceptance are well explained in chapter 1.1.6. In all cases however, sound evidence must exist for comparative estimates of DSp and DSe based on a small select panel of well-characterised samples containing the targeted analyte.

Ideally, for both comparison with a standard method or provisional recognition, a panel of, for example, 60 samples could be assembled to ensure sufficient sample size for statistical analysis of the resulting data. This would include 30 ‘true’ negatives and 30 ‘true’ positives. Wherever possible, the positives should reflect the range of analyte concentrations or activities expected in the target population. As mentioned above, each sample in this panel must represent an individual animal. Consult Chapter 2.2.5 for statistical approaches to determining methods comparability using diagnostic samples.

2. Biological modifications (Chapter 1.1.6, Section B.5.2.2)

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 reagents themselves or a change to a different type of specimen which contains the same analyte as targeted in the original validated assay (e.g. from serum to saliva). At the very least, all of the analytical criteria of the validation pathway must be re-assessed before proceeding. If the analytical requisites are met, the remaining question relates to whether or not a full diagnostic validation is required. A similar approach to the above using a panel of 60 individual reference samples may be considered. However, in this case the original test method would be considered as the standard (independent) test and the modified method would be considered

the candidate. Consult Chapter 2.2.5 for statistical approaches to determining methods comparability using diagnostic samples.

E. GROUP E

Reference animals and reference samples in this Group E are well described in chapter 1.1.6, Section B.2.1). However, there are a few points that are worth re-iterating here.

1. 'Gold standard'[2] – diagnostic specificity and diagnostic sensitivity (Chapter 1.1.6, Section B.2.1)

For conventional estimates of DSp, negative reference samples refer to true negative samples, from animals that have had no possible infection or exposure to the agent. In some situations, where the disease has never been reported in a country or limited to certain regions of a country, identification of true negative reference samples is usually not a problem. However, where the disease is endemic, samples such as these may be difficult to locate. It is often possible to obtain these samples from regions within a large country or perhaps different countries where the disease in question does not occur or has been eradicated.

For conventional estimates of DSe, positive reference samples refer to true positives. Care must be taken to ensure that the sample population is representative of the population that will be the target of the validated assay. It is generally problematic to find sufficient numbers of true positive reference animals, as determined by isolation of the organism. It may be necessary to resort to samples from animals that have been tested by a combination of methods that unequivocally classify animals as infected/exposed as discussed in chapter 1.1.6.

All samples, irrespective of origin, must be documented as they would for any other reference sample to unequivocally classify animals as infected or exposed, dependent on the fitness for purpose and proposed use of the test. As mentioned in Section A, and summarised in Figure 2, of this chapter, all reference samples should be well characterised and data documented to ensure appropriate sample selection for intended purpose.

Particularly relevant to these reference samples, the tests that are used to determine their so called 'true' disease/infection status need to be well documented in order to assess potential errors in estimates that may be carried over into the estimates for the candidate assay. Indeed, when using imperfect standard assays to define reference animal or sample status, the DSe and DSp performance estimates of the candidate assay may be flawed and often overestimated. Consult Chapter 2.2.5 for statistical considerations. 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).

F. GROUP F

1. Animals of unknown status – diagnostic specificity and diagnostic sensitivity (Chapter 1.1.6, Section B.2.2)

Latent-class models are introduced in chapter 1.1.6. They do not rely on the assumption of a perfect reference (standard or independent) test but rather estimate the accuracy of the candidate test and the reference standard with the combined test results.

Reference populations, not individual reference samples, used in latent-class studies need to be well described as summarised in Figure 2. Wherever possible, the phase of infection in the populations should be noted with respect to morbidity or mortality events, recovery, etc.

As a special note, if latent class models are to be used to ascribe estimates of DSe and DSp and include multiple laboratories in the design, it is possible to incorporate an assessment of reproducibility into the assessment. Bayesian analysis of latent class models are complex and require adherence to critical assumptions. Statistical

[2] The term "Gold Standard" is limited to a perfect reference standard as described in chapter 1.1.6, Section B.2.1.2, and Chapter 2.2.5 *Statistical approaches to validation*, Introduction and Figure 1.

assistance should be sought to help guide the analysis and describe the sampling from the target population(s), the characteristics of other tests included in the analysis, the appropriate choice of model and the estimation methods (based on peer-reviewed literature). See chapter 2.2.5 for details and Cheung *et al.,* 2021.

FURTHER READING

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.

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

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

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

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

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

*
* *

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 2014. MOST RECENT UPDATES ADOPTED IN 2024.