Rev. sci. tech. Off. int. Epiz., 1998, **17** (2), 469-486

Validation of serological assays for diagnosis of infectious diseases

R.H. Jacobson

Diagnostic Laboratory, College of Veterinary Medicine, Cornell University, Ithaca, NY 14852-5786, United States of America

Summary

Assay validation is a series of the following interrelated processes:

– an experimental process: reagents and protocols are optimised by experimentation to detect the analyte with accuracy and precision, and to ensure repeatability and reproducibility in the assay

– a relative process: its diagnostic sensitivity and diagnostic specificity are calculated relative to test results obtained from reference animal populations of known infection/exposure status

– a conditional process: classification of animals in the target population as infected or uninfected is conditional upon how well the reference animal population used to validate the assay represents the population to which the assay will be applied (accurate predictions of the infection status of animals from test results and predictive values of positive and negative test results are conditional upon the estimated prevalence of disease/infection in the target population)

– an incremental process: confidence in the validity of an assay increases over time when use confirms that it is robust as demonstrated by accurate and precise results (the assay may also achieve increasing levels of validity as it is upgraded and extended by adding reference populations of known infection status)

– a continuous process: the assay remains valid only insofar as the assay continues to provide accurate and precise results as proved through statistical verification.

Therefore, validation of diagnostic assays for infectious diseases does not end with a time-limited series of experiments based on a few reference samples. Rather, it is a process that also requires constant vigilance and maintenance, along with reassessment of its performance characteristics for each population of animals to which it is applied.

It is certain that the current movement to develop and implement accreditation criteria for veterinary diagnostic laboratories may be of little worth unless there is some assurance that the assays conducted in such laboratories are properly validated. Fully accredited laboratories may generate highly reproducible test results, but the results may still misclassify animals as to their infection status due to an improper assay validation process. Therefore, assay validation is foundational to the core product of veterinary diagnostic laboratories – test results and their interpretation.

Keywords

Animal health – Assays – Assay validation – Diagnostic techniques – Evaluation – Infectious diseases – Laboratories – Serology – Standardisation.

Introduction

What constitutes a 'validated assay'? A serological assay is considered validated if it produces test results that identify the presence or absence of a substance in serum at a specified level of statistical confidence. Inferences from test results can then be made about the infection status of animals. Examples of substances that may be detected in serum are antibodies (polyclonal or isotypic), organisms, antigens (complex or a few epitopes), nucleic acids and non-antigenic compounds; these substances are collectively termed 'analytes'. Attempts to carefully validate a serological assay for an infectious disease quickly reveal that the specific criteria required for assay validation are elusive and that the process leading to a validated assay is not standardised.

Before validation begins, a method is chosen to target a specific component in the sample that is most relevant diagnostically. Selection of a method requires thorough knowledge of it, understanding of the infectious agent and the host immune response to the agent and preliminary evidence from pilot studies that the method can succeed. Careful attention to selection of an appropriate method is essential to achieving a validated assay.

By considering the variables that affect the performance of an assay, the criteria that must be addressed in assay validation become clearer. The variables can be grouped into three categories, as follows:

– the sample: host/organism interactions affecting the analyte composition and concentration in the serum sample

– the assay system: physical, chemical, biological and technician-related factors affecting the capacity of the assay to detect a specific analyte in the sample

– the test result: the capacity of a test result, derived from the assay system, to accurately predict the status of the host relative to the analyte in question.

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

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

Factors that influence the capacity of the test result to accurately infer the infection or analyte status of the host are diagnostic sensitivity (DSn), diagnostic specificity (DSp) and prevalence of the disease in the population targeted by the assay. In this paper, the terms 'positive' and 'negative' have been reserved for test results and never refer to infection or antibody/antigen status of the host. Whenever reference is made to 'infection' or 'analyte', any method of exposure to an infectious agent that could be detected directly (e.g., antigen) or indirectly (e.g., antibody) by an assay should be inferred. DSn and DSp are derived from test results on samples obtained from selected reference animals. The degree to which the reference animals represent all of the host and environmental variables in the population targeted by the assay has a major impact on the accuracy of test result interpretation. For example, experienced diagnosticians are aware that an assay which has been validated using samples from northern European cattle may not give valid results for the distinctive populations of cattle in Africa.

The capacity of a positive or negative test result to accurately predict the infection status of the animal is a key objective of assay validation. This capacity is not only dependent upon a highly precise and accurate assay and carefully derived estimates of DSn and DSp, but is also strongly influenced by prevalence of the infection in the targeted population. Without a current estimate of the disease prevalence in that population, the interpretation of a positive or negative test result will be compromised.

Obviously, many variables must be addressed before an assay can be considered 'validated.' However, there is no consensus on whether the concept of assay validation is a time-limited process during which only the factors intrinsic to the assay are optimised and standardised or whether it includes an ongoing assessment of assay performance for as long as the assay is used. Hence, the term 'validated assay' elicits various interpretations among laboratory diagnosticians and veterinary clinicians. Therefore, a working definition of assay validation is offered as a context for the methods outlined below.

Definition of assay validation

A validated assay consistently provides test results that identify animals as being positive or negative for an analyte or process (e.g., antibody, antigen or induration at skin test site) and, by inference, accurately predicts the infection status of animals with a predetermined degree of statistical certainty. This paper will focus on the principles underlying development and maintenance of a validated assay.

The process of assay validation

Development and validation of an assay is an incremental process consisting of two principal parts. The first part is to

establish parameters and characteristics of the assay through the following methods:

a) determination of the feasibility of the method

b) development of the assay through choice, optimisation and standardisation of reagents and protocols, and

c) determination of the performance characteristics of the assay.

The second part, to assure constant validity of test results and enhancing assay validation criteria, requires the following two processes:

a) continuous monitoring of assay performance to assure that the status of 'validated assay' is merited, and

b) maintenance and enhancement of validation criteria during routine use of the assay (Fig. 1) (17).

Although some scientists may question the relevance of the second part of the process of assay validation, it is included here because an assay can be considered valid only to the extent that test results are valid, i.e., that they fall within statistically defined limits and provide accurate inferences about infection or antigen exposure status of an animal. An indirect enzyme-linked immunosorbent assay (ELISA) for detection of antibody will be used to illustrate the principles of assay validation. This is a test format that can be difficult to validate because of signal amplification of both specific and non-specific components. This methodology highlights the problems that need to be addressed in any serological assay

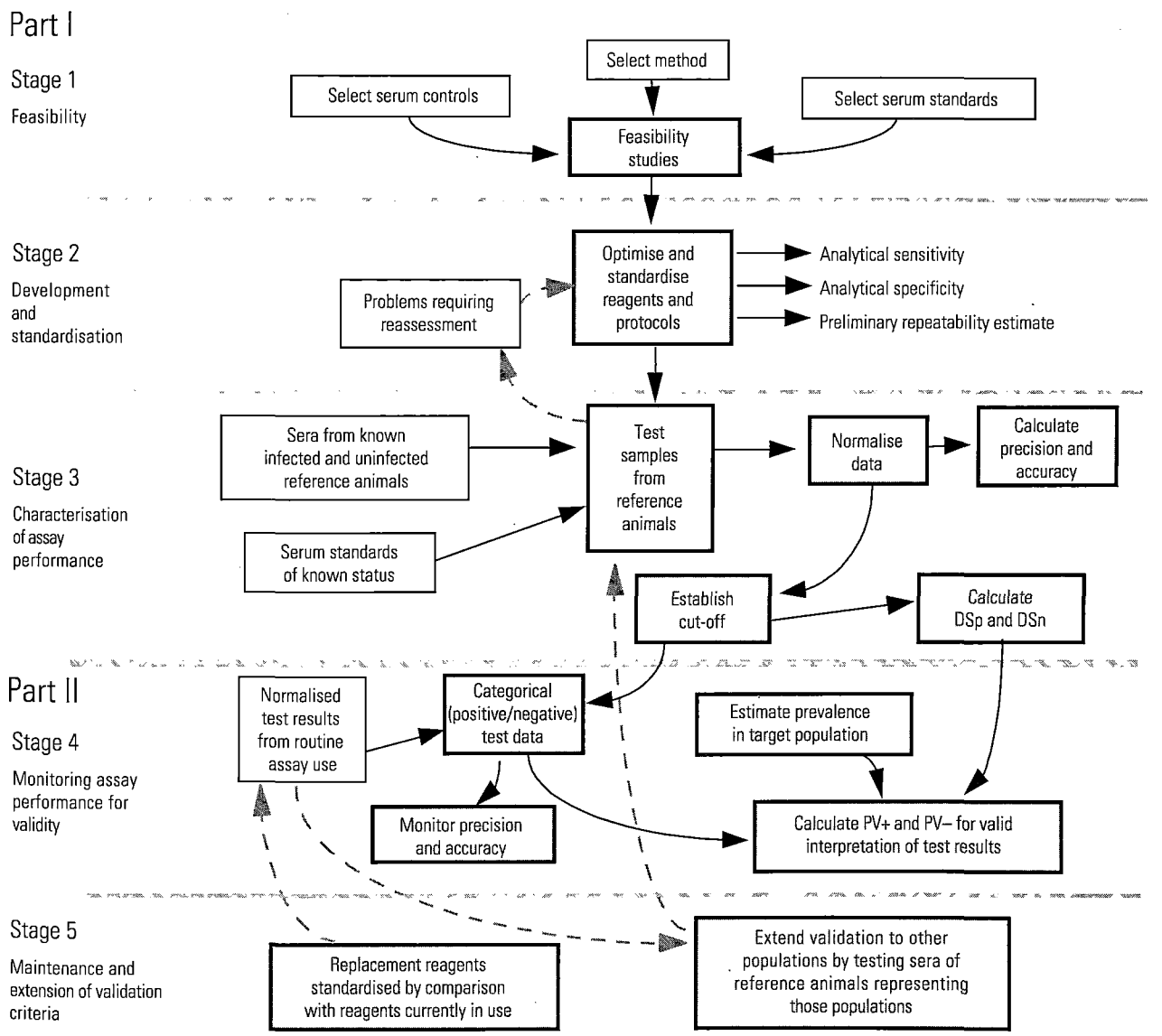

Fig. 1
The five stages in the incremental process of assay validation
Shaded boxes indicate action points within each stage of the process

validation process. The same principles are used in validation of other complex or simple assay formats.

The process of validating an assay is the responsibility of researchers and diagnosticians. The initial development and optimisation of an assay by a researcher may require further characterisation of the performance of the assay by laboratory diagnosticians before implementation. The laboratory that provides test results should have assurances, either from the literature or from research performed in that laboratory, that the assay is valid; ultimately, the laboratory that provides test results is responsible for assuring that the test results were derived from a validated assay.

First part of the process: establishing parameters and characterisation of assay performance

Feasibility studies

Feasibility studies are first performed to determine whether the selected reagents and protocol have the capacity to distinguish between a range of antibody concentrations to an infectious agent while providing minimal background activity. Such studies also give initial estimates of repeatability, analytical sensitivity and analytical specificity.

Samples for feasibility studies: serum controls

It is useful to select four or five samples (serum in our example) that range from high to low levels of antibodies against the infectious agent in question, and a sample containing no antibody. These samples will be used firstly to optimise the assay reagents and protocol, and later as serum control samples during routine runs of the assay. The samples should ideally represent known infected and uninfected animals from the population that eventually will become the target of the validated assay. The samples are preferably derived from individual animals but they may represent pools of samples from several animals. A good practice is to prepare a large volume (e.g., 10 ml) of each sample and divide it into 0.1 ml aliquots for storage at –20°C. One aliquot of each is thawed, used for experiments and held at 4°C between experiments until depleted. Then, another is thawed for further experimentation. This method provides the same source of sera with the same number of freeze/thaw cycles for all experiments (repeated freezing and thawing of serum could denature antibodies so should be avoided). Also, variation is reduced when the technician uses identical sera for all experiments rather than switching between various sera between experiments. The approach has the added advantage of generating a data trail for the repeatedly run samples. After the initial stages of assay validation are completed, one or more of the samples can become the serum control(s) that are the basis for data expression and repeatability assessments both within and between runs of the assay. They may also serve as standards if their activity has been pre-determined; such standards provide assurance that runs of the assay are producing accurate data (21).

Selection of method to achieve normalised test results

Normalisation adjusts the raw test results of all samples relative to values of controls included in each run of the assay (not to be confused with transformation of data to achieve a 'normal' [Gaussian] distribution). The method of normalisation and expression of data should be selected preferably no later than at the end of the feasibility studies. Comparisons of results from day to day and between laboratories are most accurate when normalised data are used. For example, in ELISA systems, raw optical density (absorbance) values are absolute measurements that are influenced by ambient temperatures, test parameters and photometric instrumentation. To account for this variability, results are expressed as a function of the reactivity of one or more serum control samples that are included in each run of the assay. Such data are said to be normalised or indexed to the control(s).

Data normalisation is accomplished in the indirect ELISA by expressing absorbance values in one of several ways (21). A simple and useful method is to express all optical density values as a percentage of a single positive serum control that is included on each plate. This method is adequate for most applications. A more rigorous method is to calculate results from a standard curve generated by several serum controls. This requires a more sophisticated algorithm, such as linear regression or log-logit analysis (20). This approach is more precise because it does not rely on only one control sample for data normalisation, but utilises several serum controls, adjusted to expected values, to plot a standard curve from which the sample value is extrapolated. It also allows for exclusion of a control value that may fall outside the expected confidence limits when generating the standard curve.

For assays such as virus neutralisation which are end-pointed by sample titration, each run of the assay is accepted or rejected depending on whether control values fall within predetermined limits. As sample values are not usually adjusted to a control value, the data are not normalised by the strict definition of the term.

Whatever method is used for normalisation of the data, it is essential to include additional controls for any reagent that may introduce variability and thus undermine attempts to achieve a validated assay. The normalised values for those controls need to fall within predetermined limits (e.g., within ±2 or ±3 standard deviations (SD) from the mean of many runs of each control sample).

Development and standardisation

Determination of optimal reagent concentrations and protocol parameters

Assay development follows successful pilot studies that indicate that the method has promise. It begins with optimisation of concentrations/dilutions of the antigen adsorbed to the plate, serum, enzyme-antibody conjugate and substrate solution, which are determined through checkerboard titrations of each reagent against all other reagents after confirming the best choice of reaction vessels. The process usually includes the evaluation of two or three types of microtitre plates, each with its unique binding characteristics, to minimise background activity while achieving the maximum spread in activity between negative and high positive samples. Additional experiments determine the optimal temporal, chemical and physical variables in the protocol, including incubation temperatures and durations; the type, pH and molarity of diluent, washing and blocking buffers; and equipment used in each step of the assay (for instance, pipettors and washers that give the best reproducibility). The literature is replete with papers and monographs detailing the reagents and protocols that are available for assay development (for ELISA, see 2, 14, 20).

Optimisation of the reagents and protocol should include an assessment of accuracy by inclusion of one or more serum standards which have a known level of activity for the analyte in question. An optimised assay that repeatedly achieves the same results for a serum standard and the serum controls may be designated as a standardised assay.

Repeatability: preliminary estimates

Preliminary evidence of repeatability (agreement between replicates within and between runs of the assay) is necessary to warrant further assay development. This is accomplished by evaluating results from replicates of all samples within each plate (intraplate variation), and by using the same samples run in different plates within a run and between runs of the assay (interplate variation). For ELISA, raw absorbance values are usually used at this stage of validation because it is uncertain whether the results of the high positive control serum, which could be used for calculating normalised values, are reproducible in early runs of the assay format. Also, mean values from repeated runs on each control (expected values for the controls) would not yet have been established. Three to four replicates of each control sample, run in at least five plates on five separate occasions, are sufficient to provide preliminary estimates of repeatability. Coefficients of variation (SD of replicates divided by mean of replicates), generally with values less than 20% for raw absorbance values, indicate adequate repeatability at this stage of assay development. However, if evidence of excessive variation (> 30%) is apparent for the majority of samples within and/or between runs of the assay, more preliminary studies should be conducted to determine whether stabilisation of the assay is possible or whether the test format should be abandoned. This is important because an assay that is inherently variable has a high probability of not withstanding the rigours of day-to-day testing on samples from the targeted population of animals.

Determination of analytical sensitivity and specificity

The analytical sensitivity of the assay is the smallest detectable amount of the analyte in question, and analytical specificity is the degree to which the assay does not cross-react with other analytes. These parameters are distinguished from DSn and DSp as defined below. The relative analytical sensitivity of ELISA versus immunofluorescence assay (IFA), for example, can be assessed by end-point dilution analysis which indicates the dilution of serum in which antibody is no longer detected. A quantitive estimate of analytical sensitivity can be determined by end-point titration of a sample of known antibody concentration (mg/ml). Analytical specificity is assessed by use of a panel of sera derived from animals that have experienced related infections that may stimulate cross-reactive antibodies. If the assay does not detect antibody in limiting dilutions of serum with the same efficiency as other assays, or cross-reactivity with antibodies elicited by closely related agents is commonly observed, the reagents need to be recalibrated, replaced, or the assay abandoned.

Determining assay performance characteristics

Diagnostic sensitivity and specificity

Estimates of DSn and DSp are among the primary parameters obtained during validation of an assay, and form the basis for calculation of other parameters from which inferences are made about test results. Ideally, DSn and DSp are derived from testing a series of reference samples from reference animals having known history and infection status relative to the disease/infection in question.

Diagnostic sensitivity is the proportion of known infected reference animals that give positive results in the assay; infected animals that give negative results are considered to yield false negative (FN) results. Diagnostic specificity is the proportion of uninfected reference animals that yield negative results in the assay; uninfected reference animals that give positive results are considered to yield false positive (FP) results. The number and source of reference samples used to derive DSn and DSp are of paramount importance for proper assay validation.

Size of reference serum panel required for calculations of diagnostic sensitivity and specificity

Theoretically, the number of reference samples from animals of known infection/exposure status can be calculated for determinations of DSn and DSp within statistically defined limits (5). Some assumptions must be made. A modest diagnostic performance for the assay, for example, 92% DSn and 90% DSp, should be estimated. It is better to underestimate rather than overestimate assay performance because the number of reference samples required is inversely

related to estimates of DSn and DSp (as long as these estimates do not fall below 50%). Hence, high estimates of these parameters will lead to calculation of inadequate sample sizes. The following calculations assume that the reference animals from which serum samples are acquired are a random sample from either known infected or known uninfected animals in the target population.

Number of infected reference animals required

The number of infected reference animals required to achieve an anticipated diagnostic sensitivity (± allowable error) can be approximated by the formula:

$$n = \frac{(\mathrm{DSn})(1 - \mathrm{DSn})(c)^2}{e^2}$$

where '*n*' is the number of known infected animals, 'DSn' is the worst-case assumption of the diagnostic sensitivity (i.e., the expected proportion of infected animals in the target population that will give positive test results), '*e*' is the percentage of error (expressed as a decimal) allowed in the estimate of diagnostic sensitivity, and '*c*' is the confidence interval for the estimate (modified from the Office International des Epizooties (OIE) *Manual of Standards for Diagnostic Tests and Vaccines* [17]). At a diagnostic sensitivity of 92% (±2% error allowed), with a 95% confidence (1.96 representing ±2 SDs) that the estimate is correct, the theoretical number of animals required is:

$$n = \frac{(0.92)(1 - 0.92)(1.96)^2}{(0.02)^2} = 707.$$

Table I invokes this formula to provide the theoretical number of reference animals required for various estimates of DSn and DSp at different confidence intervals, with a 2% error accepted for the estimates. If a different level of error in the estimate of DSn or DSp is allowed, the number of samples listed in the body of the Table can be multiplied by one of the factors listed in the footnote of Table I. For instance, instead of 707 samples required at a 95% confidence interval for a DSn of 92% having a 0.02 error, if 0.04 error is acceptable then the number of samples required is 177 (707 × 0.25).

The selection of 707 infected animals may be adequate to achieve reasonable estimates of DSn and DSp, provided careful sampling is performed to include as many as possible of those variables that have an impact on antibody production. A few examples of these variables are breed, age, sex, nutritional status, pregnancy, stage of infection, differing responses of individuals to infectious agents and differing host responses in chronic versus peracute infections. Also, antibody to closely related infectious agents may cause cross-reactions in the assay; if these agents occur only in one portion of the total population targeted by the assay, but are not represented in the panel of reference sera, then estimates of DSn and DSp will be errant. It is desirable, therefore, to increase the sample size to approximately 1,000 samples from infected reference animals. Although this number of sera may be difficult to obtain, it should be the ultimate goal as outlined below.

Increasing the expected DSn for the new test to 99% would decrease the theoretical number of animals required to only 95 (Table I; see 95% confidence level). This estimate is inadequate because it is impossible to fully represent all variables found in a target population of 25 million animals,

Table I
Theoretical number of samples from animals of known infection status required for validating an assay

Estimated DSn or DSp (%)	Confidence levels 75% (1.0694)	80% (1.2814)	85% (1.4532)	90% (1.6462)	95% (1.9599)	99% (2.5758)
80%	457	657	845	1,084	1,536	2,654
82%	422	606	779	1,000	1,417	2,448
84%	384	552	710	911	1,291	2,229
86%	344	494	636	816	1,156	1,997
88%	302	433	558	715	1,014	1,752
90%	257	369	475	610	864	1,493
92%	210	302	389	499	**707**	1,221
94%	161	232	298	382	542	935
95%	136	195	251	322	456	788
96%	110	158	203	260	369	637
97%	83	119	154	197	279	483
98%	56	80	103	133	188	325
99%	28	41	52	67	95	164

Percent error allowed in the estimate of DSn or DSp = 0.02. To determine the number of samples required for 0.01 allowable error, multiply number of samples in Table by a factor of 4; for 0.03 error, a factor of 0.444; for 0.04 error, a factor of 0.25; for 0.05 error, multiply by a factor of 0.16

DSn : diagnostic sensitivity
DSp : diagnostic specificity

for example, using a sample of only 95 animals even if they are derived from the target population. These calculations of sample numbers assume a normal distribution of values for each of an indeterminate number of continuous variables that may affect antibody production in the target population. As it is unlikely that the assumptions of normality are true under these circumstances, particularly when the sample size is small, it is recommended that a minimum of approximately 300 samples are tested to provide added confidence in the estimates of DSn and DSp.

Number of uninfected reference animals required

As estimated rates are being used here, the same formula is theoretically relevant for calculating the number of known uninfected reference animals to estimate the DSp (the rate of negative test results among known uninfected animals) for the new assay. Again, the desired rate (DSp in this case) is inversely related to the number of samples required to achieve a precise estimate of that DSp. Therefore, despite the fact that high DSp is usually desired to minimise FP test results in the target population, it is important to select a low estimate of DSp rather than a high one for the eventual validated assay. The lower estimate will assure a sufficient sample of uninfected animals to provide confidence in the estimate of DSp, should the need arise to assign a high DSn in the assay (with a commensurate reduction in DSp). If it is estimated that the new assay will achieve a DSp of 90%, the calculated number of animals required is 864 (at the 95% confidence level). Many more biological variables may contribute to FP results (e.g., cross-reactive antibodies to many other agents) than to FN results (for most but not all pathogens, animals generally develop antibody responses and thus are not falsely negative). It is necessary, therefore, to account for this probable increased variance that would affect the estimate of DSp. This suggests that testing from 1,000 to 5,000 known uninfected animals would be a laudable goal to assure a very high level of confidence in the estimate of DSp. It is recognised, however, that such numbers of reference animals may be unrealistic (see Section on 'Alternative sources and numbers of reference sera' below for resolution of this problem).

Intended use of the assay: effect on number of samples required

The intended use of the assay may affect decisions about the number of samples required to establish DSn and DSp for the new assay. Screening, confirmatory or 'all-purpose' diagnostic assays require different approaches for establishing the assay characteristics. When a screening assay is needed for detection of a pathogenic disease such as foot and mouth disease, it is necessary to reduce the likelihood that infected animals will be misclassified as uninfected. Accordingly, the percentage of error allowed in the estimate of DSn ('*e*' in the formula given above) must be minimised. Alternatively, when an assay is designed for a less pathogenic disease, a high DSp is selected, with a commensurate increase in the number of FN results; this will reduce the likelihood that uninfected animals will be classified as infected. To optimise the DSp of the assay, large numbers of uninfected reference animals should be evaluated to minimise sampling errors. Assays with high specificity are often used as confirmatory assays. An all-purpose diagnostic assay may place the cut-off in the centre of the FP-FN range (see 'Selection of a cut-off [positive/negative threshold]' below). If the assay is intended for use on sera from vaccinated animals, separate estimates of DSp and DSn may be required for vaccinated versus non-vaccinated animals to properly reflect the impact of vaccination on test interpretation.

Alternative sources and numbers of reference sera

It is very difficult, if not impossible, to find a large number of proven uninfected animals from the target population where the disease/infection is endemic or where vaccination is not commonly used. Therefore, it may be necessary to start stage 3 (Fig. 1) of the validation process with small panels of sera. When the assay is used routinely, confirmatory data should be obtained whenever possible to update estimates of DSn and DSp. There is a very high risk that the assay will not be accurate when only a few reference animals are used as a basis for validation.

In some situations, it is necessary to begin validation studies using animals located in a geographically distinct region in which the infection in question does not exist. Assembling a panel of sera from known infected animals may be equally as difficult. Of necessity, these reference animals may be from a region removed geographically from the target population or may even be from another continent. Results of tests on these animals serve only as a starting point toward establishing estimates of DSn and DSp for the target population. As samples from animals in the target population are subsequently tested and several thousand results are acquired, it is then possible to estimate a reasonable cut-off for the assay through some of the newer statistical techniques, such as mixture analysis and cluster analysis (3). A discussion of this methodology is included in the Section entitled 'Intrinsic cut-off established where no reference animals are available' below.

Standards of comparison: the basis for defining certain assay performance characteristics

In serology, the term 'gold standard' or 'benchmark' refers to the results of a method or combination of methods that are regarded to classify animals as infected or not infected. It may also refer to a method that classifies samples as positive or negative, such as another seroassay system. Accordingly, the so-called gold standard carries several connotations and may not be as perfect as the term implies; indeed, the gold standard result may be equivocal relative to the infection status of the animal. Therefore, in this paper the term 'gold standard' is supplanted by various 'standards of comparison' as the basis for defining certain performance characteristics of the new assay.

The results of the new assay are deemed correct or incorrect relative to the standard of comparison. Several methods have been described which can be used with varying success to characterise the infection status of animals that serve as a source of reference sera.

Verification of infection: an absolute standard of comparison

If an infectious agent or definitive histopathological criterion is detected, this usually constitutes an unequivocal standard of comparison that is legitimately called a gold standard for classifying the animal as infected. However, even this standard has limitations. Reference animals with gold-standard proof of infection may already have generated strong immune responses and may therefore possess easily detected antibody. In contrast, the target population for the new assay may consist of many animals that have early infections or latent infections that are not accompanied by detectable antibody responses in analytically insensitive tests, or would not be detected by culture or histopathology. Therefore, using only reference animals that have confirmatory culture or histopathology may produce higher estimates of DSn than are realistic for the target population. So, even an unequivocal standard that classifies animals as infected may have limitations as a basis of comparison for the new assay.

Comparative serology: a relative standard of comparison

To obtain definitive proof of infection through culture or isolation techniques may be impractical, technically difficult or impossible. Therefore, other methods must serve as the standard of comparison for the new assay. If other assays have acceptable and established performance characteristics, such as the Rose Bengal screening test followed by the complement fixation confirmatory test for detection of antibody to *Brucella abortus*, then the collective results of these assays provide a useful composite-based standard to which the new assay may be compared. When the new assay is evaluated by comparison with another serological assay or combination of assays, the estimates of DSn and DSp for the new assay are called relative diagnostic sensitivity and relative diagnostic specificity. These standards of comparison, however, have their own established levels of false positivity and false negativity which are sources of error that will be compounded in calculations of DSn and DSp of the new assay. It follows that the greater the rate of false positivity or false negativity in the assay that is used as the standard of comparison, the more the performance characteristics of the new assay will be undermined.

It is possible that the new assay has a greater sensitivity and/or specificity than the assay(s) used as the standard of comparison. This is suspected when the new assay gives a higher percentage of false positive or false negative results than expected. One method to assess this scenario is to first use mixture or cluster analysis as described below (see Section entitled 'Intrinsic cut-off established when no reference animals are available'; reference 3) to select a tentative cut-off for the new assay. The samples are classified as positive or negative based on that cut-off. The roles of the two tests are then reversed, making the new test the standard of comparison (independent variable). The results may infer that the new assay has better performance characteristics than the established assays. The estimated performance characteristics should be confirmed by additional studies that evaluate sera from animals of known infection status or sera from experimentally infected animals.

Experimental infection or vaccination: an adjunct standard of comparison

Another standard for assessment of antibody responses is sera obtained sequentially over several months from each of several experimentally infected or vaccinated animals. These sera should reveal the ability of the assay to detect early antibody production and the kinetics of antibody production to the agent in question. If it is evident that animals become infected, shed organisms in low numbers, but have no detectable antibody in the new assay during the first two to three months of infection, the analytical sensitivity of the assay may be inadequate and estimates of diagnostic sensitivity will be low. Alternatively, if antibody appears quickly after inoculation of the infectious agent (and earlier than in the conventional assays that are used as standards of comparison) the new assay may have greater analytical sensitivity and associated diagnostic sensitivity than the conventional assay. Experimental infections may also provide evidence of class-specific antibody responses. This is useful for selecting reagents that will detect early (IgM) responses, or other antibody classes appropriate to the agent such as IgE for helminth infections.

Caution must be exercised when interpreting the antibody responses of experimentally induced infections. The strain of cultured organism, route of exposure and dose are just three of the variables that may elicit antibody responses that are quantitatively and qualitatively atypical of natural infection in the target population. The same is true of vaccination. Therefore, it is essential that experimentally induced antibody responses are relevant to those occurring in natural outbreaks of disease caused by the same infectious agent, or the estimates of relative DSn and DSp may be in error. Due to the difficulty of achieving equivalent responses from naturally infected and experimentally infected/vaccinated animals, the relative DSn and DSp data derived from such animals should be considered as an adjunct standard of comparison, and should not be used alone to determine the relative DSn and DSp of the new assay.

Verification of uninfected/unexposed status: a composite standard

Classification of animals as unexposed to the agent in question with absolute certainty is not possible. Ante-mortem tests cannot rule out the possibility of FN results. A combination of several sources of information may help to determine whether it is probable that the reference animals have not been exposed to the agent in question. Ideally, reference animals chosen to represent the unexposed group for assessment of DSp are selected from the following:

a) geographical areas within the target population where the disease has not been endemic for a period of approximately three years (this interval may be longer or shorter, depending upon the particular disease)

b) herds from those areas that have displayed no clinical signs of the disease for at least three years, and have not been vaccinated against the agent in question

c) herds closed to importation of animals from endemic areas and with no infected neighbouring herds, and

d) herds with no evidence of antibody to the agent in question based on repeated testing over the past two to three years.

If all of these criteria are met, it is reasonably certain that these animals have had no exposure to the agent in question. Such animals could then be used as a source of reference sera for the unexposed reference animal group.

Precision, repeatability, reproducibility and accuracy

Repeatability and reproducibility are estimates of precision in the assay. Precision is a measure of dispersion of results for a repeatedly tested sample; a small amount of dispersion indicates a precise assay. Repeatability has two elements: the amount of agreement between two or three replicates of each sample within a run of the assay, and the amount of between-run agreement for the normalised values of each control sample. Reproducibility is the amount of agreement between results of samples tested in different laboratories. Accuracy is the amount of agreement between a test value and the expected value for an analyte in a standard sample of known activity (e.g., titre or concentration). An accurate assay will have a minimum of bias and random error. An assay system may be precise but not accurate if the test results do not agree with the expected values of the standard, but it cannot be accurate if it is not precise.

Evaluation of repeatability

The preliminary evidence of repeatability (as described above) was based on the use of raw data. Large coefficients of variation (CVs) with values approaching 20%-30% were thus acceptable. The main body of repeatability data is obtained from normalised (not raw) data so the acceptable range of CVs will be lower. To determine repeatability and accuracy, it is convenient to use normalised results from the many runs of the new assay that are required to assess the sera of reference animals. At least 10, and preferably 20 runs of the assay will give reasonable initial estimates of these parameters. For within-run repeatability, the mean ± SD is computed for replicates of each serum tested. The CVs for normalised data from the replicates of each serum should not exceed 10% unless the mean value approaches zero, in which case CVs are not meaningful. Between-run repeatability within a laboratory can be based on the normalised test results for the serum controls, representing negative, low and high antibody levels. The mean of replicates for each control sample tested in each of about 20 runs of the assay is recorded. Values are generally acceptable if they remain within ±2 SD of the mean of all runs. These values may be plotted as points on Levey-Jennings charts (1), using one chart for each control to visualise the results (Fig. 2). The lines representing ±1, ±2 and ±3 SD from

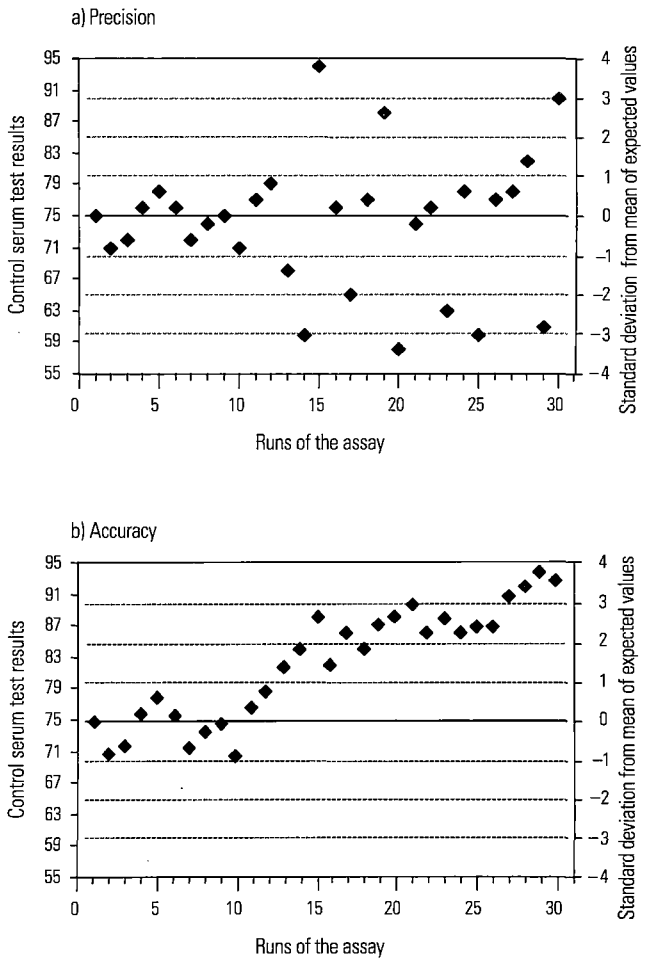

Fig. 2
Charts of control values illustrating precision (a) and accuracy (b)
Test results for each control are plotted daily on separate charts. Each tick on the x-axis is a run of the assay. After about 20 runs of the assay are completed, six horizontal lines are drawn on the each chart, representing ±1, ±2 and ±3 standard deviation above and below the mean of the 20 values for each control. Panel *a)* represents an increased dispersion (reduced precision) for test results of a serum control after the 12th run. Panel *b)* represents excellent precision but a shift toward higher test values (reduced accuracy) after the 12th run

the mean can be used as measures of dispersion (16). Precision is reduced as dispersion increases (Fig. 2a). As routine runs of the assay are eventually conducted on the target population, the charts can represent the last 30 consecutive runs of the assay; a running mean with its SD will then constitute the constantly updated chart for each sample. It may be necessary to customise decision criteria for a given assay because of inherent variation attributable to the host/pathogen system.

Accuracy can be assessed by inclusion of one or more standards in each run of the assay. A standard is defined as a sample for which the concentration or titre of the analyte has been established by methods independent of the assay being validated. The standards may be control sera, provided that the amount of analyte (e.g., titre, concentration) in each one has been previously determined by comparison with primary or secondary reference standards (21), and the control is not used in the data normalisation process. The Levey-Jennings charts may be used to assess accuracy in the assay (Fig. 2b). A rapid shift or a trend upwards or downwards in the pattern of a standard indicates that a bias has been introduced, thus reducing accuracy. The extent of the shift will suggest whether or not corrective measures need to be taken (16).

Evaluation of reproducibility

Reproducibility of the assay is determined when several laboratories using the identical assay (protocol, reagents and controls) compare results. A group of at least 10 samples (preferably duplicated to a total of 20 with encoded identifications) representing the full range of expected analyte concentrations is evaluated in each laboratory. The extent to which the collective results for each sample deviate from expected values is a measure of assay reproducibility. The evaluation is based upon values obtained in the assay (e.g., normalised data on a continuous scale) and not interpretations of those values (e.g., 'positive' or 'negative' categorical data). The degree of concordance of between-laboratory data is one more basis for determining whether the performance characteristics of the assay are adequate to constitute a validated assay. There are no universal decision criteria for gauging reproducibility. The criteria used with Levey-Jennings charts would be adequate for decisions of acceptance/rejection.

Evaluation of technician error by Levey-Jennings charts

As technician error is the greatest source of variation for most assays, it is useful to prepare separate Levey-Jennings charts representing repeatability and accuracy data for each technician. These are prepared in addition to charts representing the collective efforts of all the technicians who run the assay within a laboratory. If variation between technicians and/or between laboratories is large, then it is necessary to determine whether the assay is inherently subject to variation (i.e., lacks robustness), or whether certain technicians are incapable of obtaining repeatable results.

Selection of a cut-off (positive/negative threshold)

To achieve DSn and DSp estimates for the new assay, the test results must be reduced to positive or negative categories. Insertion of a cut-off point (threshold or decision limit) on the continuous scale of test results allows calculation of DSn and DSp. Although many methods have been described for this purpose, three examples will illustrate different approaches together with their advantages or disadvantages. The first is a cut-off based on the distribution of test results from uninfected and infected reference animals, which allows for calculation of DSn and DSp. A second approach is to establish a cut-off based only on uninfected reference animals; this provides an estimate of DSp but not DSn. The third provides an 'intrinsic cut-off' based on test results from sera drawn randomly from within the target population with no prior knowledge of the infection status of the source animals (3). No estimates of DSn and DSp are obtained by this method but these could be determined as confirmatory data are accumulated.

Cut-off based on test results of reference sera from uninfected and infected animals

The choice of a cut-off is based on the three following criteria:

- frequency distributions of normalised test results from two sets of reference samples, one from animals infected with the agent in question and the other from uninfected animals
- the prevalence of disease in the target population
- the impact of FP and FN test results (19).

Selecting a cut-off by visual inspection of frequency distributions

The frequency distributions for 600 infected and 1,400 uninfected animals (Fig. 3) indicate an overlapping region of assay results (the perfect test with no overlap, yielding 100% DSn and 100% DSp, rarely – if ever – exists). Placing the cut-off at the intersection of the two distributions results in rates of 5% FN and 4.7% FP for the assay. The extent of the overlap may vary considerably from one assay to another. Moving the cut-off to the left minimises FN results (thus favouring greater DSn) or to the right minimises FP results (thereby favouring greater DSp) depending on the intended application of the assay. This method has the advantage of being simple and flexible, and requires no statistical calculations or assumptions about the normality of the two distributions.

Selecting a cut-off by a modified receiver-operator characteristics analysis

Another useful tool for determining the performance characteristics of an assay is a receiver-operator characteristic (ROC) curve (22). ROC curves are plots of true positive (TP)

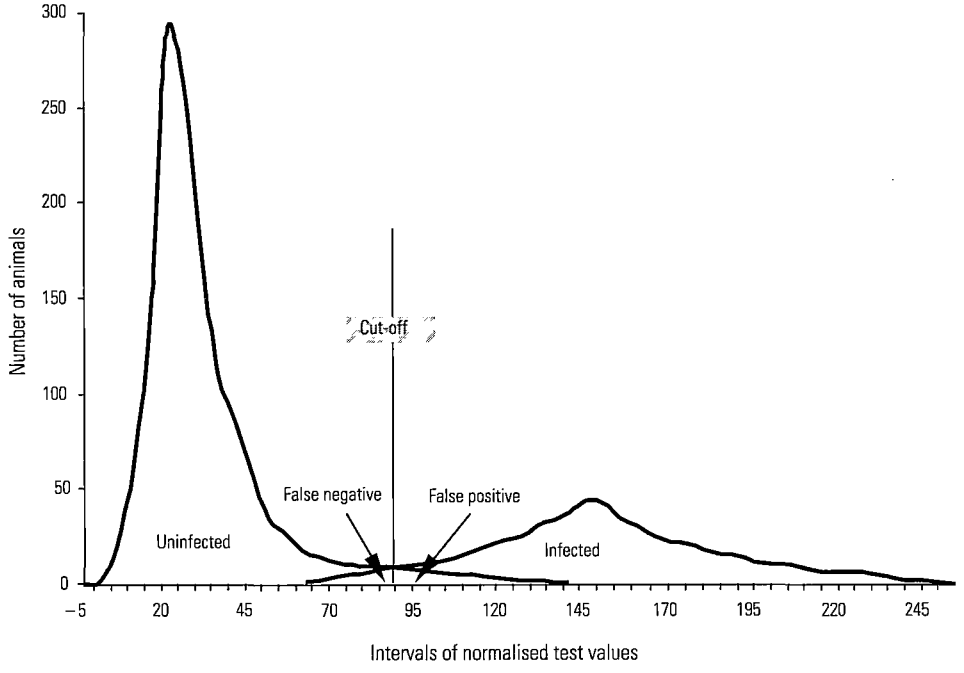

Fig. 3
Hypothetical frequency distribution of normalised test values (e.g., titre, absorbance, percent of positive control) from sera of reference animals of known infection status
A line representing the cut-off is set at the intersection of the two frequency distributions

rates (sensitivity on the y-axis) against FP rates (1 – specificity on the x-axis) using test results from serum panels of known uninfected and infected animals. The points that define the curve merely represent a series of cut-off values. When ROC curves are plotted for several assays on the same chart, the assay representing the largest area under the ROC curve is considered the most accurate. This is a simple way to compare two assays graphically for their degree of concordance. ROC curves are also useful for selecting a cut-off when the relative cost of FN and FP results can be estimated (15). Standard ROC curves are not so useful for selecting an optimal cut-off for a screening or confirmatory assay. The commonly used Kappa statistic is not recommended because of its dependence on prevalence and the possibility that concurrence between two tests can occur by chance alone if the two tests being compared both have DSn and DSp exceeding 50% (J.W. Wilesmith, personal communication).

Given the fact that it is difficult to read an optimal cut-off from an ROC curve, a modified ROC curve has been devised to make the choice of a cut-off more intuitive while remaining statistically accurate (4, 13). The modified ROC plots the TP rate (DSn) and the true negative (TN) rate (DSp) separately for each cut-off in a series of cut-offs that are represented by increasing intervals of test values on the horizontal axis (Fig. 4). Overlaying the resultant DSn and DSp curves on the frequency distributions from which they were derived illustrates the relationship between the overlap in the frequency distributions versus the DSn and DSp at various cut-offs in that overlapping region. Selecting a series of different cut-offs while moving from left to right on the horizontal axis clearly demonstrates the effect of a cut-off selection on DSn and DSp.

Cut-off based on test results from uninfected animals only

The mean of test values obtained from a large group of known uninfected animals, +2 SD or +3 SD, is often used as a cut-off in ELISA. Under the assumption of a normal (Gaussian) distribution, the expected DSp would be approximately 97.5%, 97.7%, or 99.9% if the cut-off value selected was equal to the mean of the negative reference serum plus 1.96, 2 or 3 times the SD, respectively. Given that test results, particularly those obtained from uninfected animals, are seldom normally distributed but are rather skewed to the right, errant estimates of DSp may occur. Addition of 2 or 3 SDs to the mean may not plot a cut-off that would represent the expected 97.7% or 99.9% of test results, respectively. In fact, if the values for most of the uninfected animals are minimal (thus resulting in a low mean) but the distribution has the commonly observed long tail to the right, then the cut-offs from SD calculations may result in a higher proportion of FP results than estimated by the SD statistic. Infected animals, on the other hand, often give frequency distributions that approximate a normal distribution.

A preferred alternative to the SD parametric statistic is to use a percentile of the values (e.g., 99% of the values from

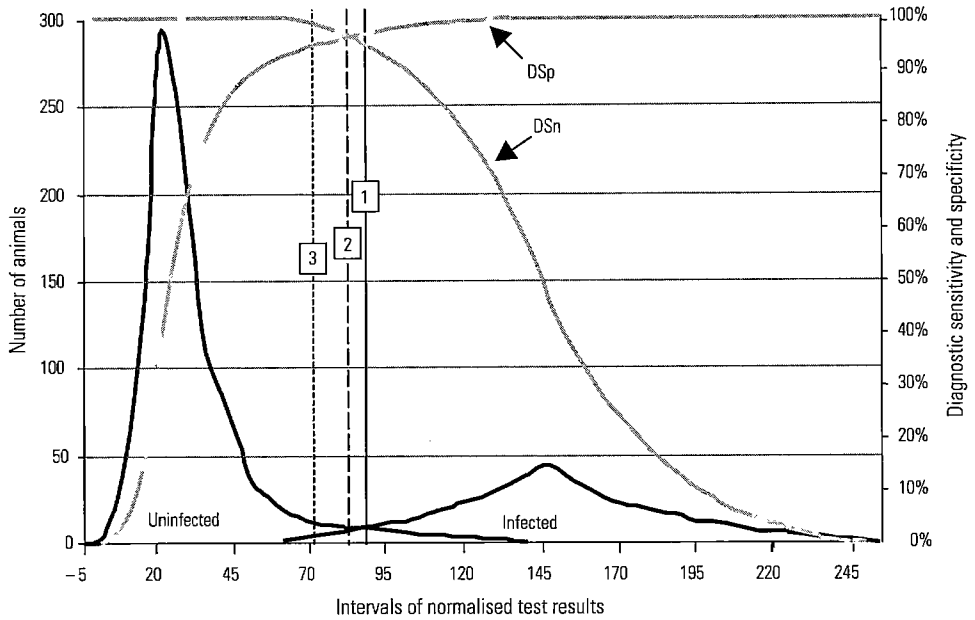

Fig. 4
Selection of cut-offs by a modified receiver-operator characteristic (ROC) analysis that plots diagnostic sensitivity and diagnostic specificity (y-axis on right-hand side of chart) as a function of cut-off (x-axis)
The ROC curves are superimposed on the frequency distributions of Figure 3. Three cut-offs are shown: number 1 represents a cut-off of 87 units on the x-axis and was chosen by visual inspection as described in Figure 3; number 2 is at 82 units where DSn and DSp are equal (97.5%); and number 3 is at 70 units, representing the greatest diagnostic accuracy (total of DSn and DSp) for the assay, 94.9% and 98.8%, respectively

uninfected animals). This approach is not subject to error associated with lack of normality in the distributions. However, using only uninfected animals does not allow calculation of DSn. This approach is suggested, therefore, only for tests where estimates of DSp are of utmost importance and DSn is of little value, such as in a confirmatory test to rule out FP test results.

Intrinsic cut-off established when no reference animals are available

For many diseases/infections, it is impossible to obtain a sufficient number of samples from known infected or uninfected animals to establish a cut-off. Also, when reference sera are not from the target population, the selected cut-off may be inappropriate for the target population. It is possible to base a cut-off solely on distribution analysis of the data from endemic animals in the target population. If a bimodal distribution clearly separates the distributions of infected versus uninfected animals, a cut-off may be selected by visual inspection of the plotted data alone. However, it is much better to have a statistical basis for selecting the cut-off. The analysis of mixture distributions has been described as a powerful approach for an unbiased estimation of sero-prevalence when sera from known uninfected controls were not available (3). The foremost reason for using sera obtained at random from the target population in establishing a cut-off is that it avoids the bias that may occur when the assumption is made that sera from a reference population are representative of the target population. The disadvantage of this approach is that DSn and DSp cannot be calculated. Only by post-test confirmation of infection status using a standard of comparison method can these parameters be established.

Multiple cut-offs: adding a 'suspicious' category to negative and positive results

If considerable overlap occurs in the distributions of test values from known infected and uninfected animals, it is difficult to select a cut-off that will accurately classify the infection status of animals. Rather than a single cut-off, two cut-offs can be selected, one that defines a high DSn (e.g., 99% of the reference sera from infected animals give results above the cut-off), and a second that defines a high DSp (e.g., 99% of the reference sera from uninfected animals give results below the cut-off). The values that fall between these percentiles would then be classified as suspicious or equivocal and would require testing by a confirmatory assay or retesting of the animal at a later time for detection of seroconversion.

Calculation of diagnostic sensitivity and specificity

The selection of a cut-off allows classification of test results into positive or negative categories. Calculations of DSn and DSp are aided by associating the positive/negative categorical data with the known status (standard of comparison) for each animal in a two-way (2 × 2) table (Fig. 5). After the cut-off has been established, results of tests on reference sera can be classified as TP or TN. These designations indicate agreement between the test results and those of the standard of comparison. Alternatively, results for reference sera are classified as FP or FN, which indicates disagreement with the

standard of comparison. Diagnostic sensitivity is calculated as TP/(TP + FN), whereas diagnostic specificity is TN/(TN + FP); the results of both calculations are usually expressed as percentages (Fig. 5).

		Reference animals of known infection status	
		Infected	Uninfected
Test result	Positive	570 (A)	46 (B)
	Negative	30 (C)	1,354 (D)
		Diagnostic sensitivity $\frac{A}{A+C} = \frac{570}{600} = 95.0\%$	Diagnostic specificity $\frac{D}{D+B} = \frac{1,354}{1,400} = 96.7\%$

Fig. 5
Calculations of diagnostic sensitivity and diagnostic specificity aided by a 2 × 2 Table that associates infection status with test results from the 600 infected and 1,400 uninfected reference animals depicted in Figure 3

Estimates of DSn and DSp are, therefore, entirely dependent upon the characteristics of the reference population; the estimates may have little relevance to the target population if animals used to obtain those estimates are not representative of that population. This is particularly true if an assay is transferred to another continent and a completely different population of animals. In that event, estimates of DSn and DSp need to be re-established for the new target population by revalidating it through subjection to stages 3 to 5 of the assay validation process (Fig. 1).

Interpretation of test results

Test results are useful only if the inferences made from them are accurate. A common error is to assume that an assay with 99% DSn and 99% DSp will generate one FP and one FN result for every 100 tests on animals from the target population. The assay may be precise and accurate yet produce test results that do not accurately predict infection status. For example, if the prevalence of disease in a target population is only one per 1,000 animals, and the FP test rate is one per 100 animals (99% DSp), then for every 1,000 tests on that population 10 will be FP and one will be TP (if the DSn is greater than 50%). Hence, only about 9% of positive test results will accurately predict the infection status of the animal; the test result will be wrong 91% of the time. This example illustrates that the positive predictive value (PV+) is not a direct correlate of DSp, but rather is a function of prevalence. So, calculations of PV+ and PV– from the test results on reference sera are irrelevant since infected and uninfected reference animals are not selected to mirror the prevalence of the target population. Rather, the estimated prevalence of the target population is the relevant prevalence figure for calculations PV+ and PV– from test results.

Estimating true prevalence from apparent prevalence

Estimation of the prevalence of infection for use in calculations of predictive values is often difficult. If the DSn and DSp are well established for an assay, a herd test using that assay will provide the apparent prevalence of infection in that herd. From these test results, the true prevalence can be estimated (18) using the following formula:

$$TP = \frac{AP + DSp - 1}{DSn + DSp - 1}$$

where TP equals estimated true prevalence and AP equals apparent prevalence (number of test positives divided by the number of samples tested).

Determining predictive values of positive and negative test results

An intuitive method for calculating predictive values for positive and negative test results is shown in Figure 6. A look-up chart (Table II) is also given to illustrate the impact of prevalence on predictive values.

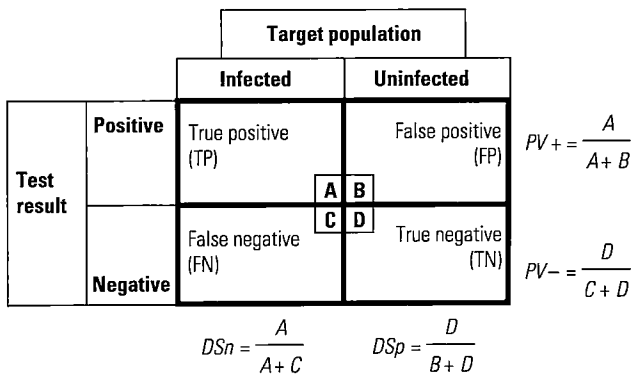

Fig. 6
Intuitive method for calculation of predictive values of positive (PV+) and negative (PV–) test results from animals in the target population

Given:
Calculations using a hypothetical group of 10,000 animals from the target population
Diagnostic sensitivity (DSn) = 99%
Diagnostic specificity (DSp) = 99%
Estimated prevalence of infection in target population = 5%

Calculations:
Percentage infected : 10,000 × 5% prevalence = 500 animals
Number of TP tests : (DSn) × (% infected) = 0.99 × 500 = 495
Number of FN tests : (% infected) – (TP) = 500 – 495 = 5
Number uninfected : 10,000 – infected = 10,000 – 500 = 9,500
Number of TN tests : (DSp) × uninfected = 0.99 × 9,500 = 9,405
Number of FP tests : Number uninfected – TN = 9,500 – 9,405 = 95

Predictive values for test results on target population:
For a positive test result : (PV+) = TP/(TP+FP) = 495/(495 + 95) = 83.9%
For a negative test result : (PV–) = TN/(TN+FN) = 9,405/(9,405 + 5) = 99.9%

Impact of infection prevalence on interpretation of test results

If the prevalence in the target population is relatively high, for example 10%, then the PV– and PV+ are 99.9% and 91.7%, respectively, for an assay that has DSn and DSp of 99% (Table II). A prevalence of 5% gives a PV– of 99.9% and a PV+ of 83.9%. However, when prevalence drops to 0.1%, for example during a disease eradication campaign, the same test

Table II

Predictive values for a positive or negative test result, expressed as a probability (%) that the test result correctly classifies the infection status of an animal

In the centre two columns of the chart, go down to the row listing the combination of the diagnostic sensitivity and specificity of the assay; then go laterally to the column representing the estimated prevalence of infection. At the intersection of the column and row is the PV+ (left panel) or PV– (right panel)

Predictive value of a positive test result (%): estimated prevalence of infection*								Assay		Predictive value of a negative test result (%): estimated prevalence of infection*							
40%	25%	10%	5%	1%	0.5%	0.1%	0.01%	Diagnostic specificity	Diagnostic sensitivity	40%	25%	10%	5%	1%	0.5%	0.1%	0.01%
100	100	100	100	100	100	100	100	100%	100%	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100%	99%	99.3	99.7	99.9	99.9	100	100	100	100
100	100	100	100	100	100	100	100	100%	98%	98.7	99.3	99.8	99.9	100	100	100	100
100	100	100	100	100	100	100	100	100%	95%	96.8	98.4	99.4	99.7	99.9	100	100	100
100	100	100	100	100	100	100	100	100%	90%	93.8	96.8	98.9	99.5	99.9	99.9	100	100
100	100	100	100	100	100	100	100	100%	80%	88.2	93.8	97.8	99.0	99.8	99.9	100	100
100	100	100	100	100	100	100	100	100%	65%	81.1	89.6	96.3	98.2	99.6	99.8	100	100
100	100	100	100	100	100	100	100	100%	50%	75.0	85.7	94.7	97.4	99.5	99.7	99.9	100
98.5	97.1	91.7	84.0	50.3	33.4	9.1	1.0	99%	100%	100	100	100	100	100	100	100	100
98.5	97.1	91.7	83.9	50.0	33.2	9.0	1.0	99%	99%	99.3	99.7	99.9	99.9	100	100	100	100
98.5	97.0	91.6	83.8	49.7	33.0	8.9	1.0	99%	98%	98.7	99.3	99.8	99.9	100	100	100	100
98.4	96.9	91.3	83.3	49.0	32.3	8.7	0.9	99%	95%	96.7	98.3	99.4	99.7	99.9	100	100	100
98.4	96.8	90.9	82.6	47.6	31.1	8.3	0.9	99%	90%	93.7	96.7	98.9	99.5	99.9	99.9	100	100
98.2	96.4	89.9	80.8	44.7	28.7	7.4	0.8	99%	80%	88.1	93.7	97.8	98.9	99.8	99.9	100	100
97.7	95.6	87.8	77.4	39.6	24.6	6.1	0.6	99%	65%	80.9	89.5	96.2	98.2	99.6	99.8	100	100
97.1	94.3	84.7	72.5	33.6	20.1	4.8	0.5	99%	50%	74.8	85.6	94.7	97.4	99.5	99.7	99.9	100
97.1	94.3	84.7	72.5	33.6	20.1	4.8	0.5	98%	100%	100	100	100	100	100	100	100	100
97.1	94.3	84.6	72.3	33.3	19.9	4.7	0.5	98%	99%	99.3	99.7	99.9	99.9	100	100	100	100
97.0	94.2	84.5	72.1	33.1	19.8	4.7	0.5	98%	98%	98.7	99.3	99.8	99.9	100	100	100	100
96.9	94.1	84.1	71.4	32.4	19.3	4.5	0.5	98%	95%	96.7	98.3	99.4	99.7	99.9	100	100	100
96.8	93.8	83.3	70.3	31.3	18.4	4.3	0.4	98%	90%	93.6	96.7	98.9	99.5	99.9	99.9	100	100
96.4	93.0	81.6	67.8	28.8	16.7	3.8	0.4	98%	80%	88.0	93.6	97.8	98.9	99.8	99.9	100	100
95.6	91.5	78.3	63.1	24.7	14.0	3.2	0.3	98%	65%	80.8	89.4	96.2	98.2	99.6	99.8	100	100
94.3	89.3	73.5	56.8	20.2	11.2	2.4	0.2	98%	50%	74.6	85.5	94.6	97.4	99.5	99.7	99.9	100
93.0	87.0	69.0	51.3	16.8	9.1	2.0	0.2	95%	100%	100	100	100	100	100	100	100	100
93.0	86.8	68.8	51.0	16.7	9.0	1.9	0.2	95%	99%	99.3	99.7	99.9	99.9	100	100	100	100
92.9	86.7	68.5	50.8	16.5	9.0	1.9	0.2	95%	98%	98.6	99.3	99.8	99.9	100	100	100	100
92.7	86.4	67.9	50.0	16.1	8.7	1.9	0.2	95%	95%	96.6	98.3	99.4	99.7	99.9	100	100	100
92.3	85.7	66.7	48.6	15.4	8.3	1.8	0.2	95%	90%	93.4	96.6	98.8	99.4	99.9	99.9	100	100
91.4	84.2	64.0	45.7	13.9	7.4	1.6	0.2	95%	80%	87.7	93.4	97.7	98.9	99.8	99.9	100	100
89.7	81.3	59.1	40.6	11.6	6.1	1.3	0.1	95%	65%	80.3	89.1	96.1	98.1	99.6	99.8	100	100
87.0	76.9	52.6	34.5	9.2	4.8	1.0	0.1	95%	50%	74.0	85.1	94.5	97.3	99.5	99.7	99.9	100
87.0	76.9	52.6	34.5	9.2	4.8	1.0	0.1	90%	100%	100	100	100	100	100	100	100	100
86.8	76.7	52.4	34.3	9.1	4.7	1.0	0.1	90%	99%	99.3	99.6	99.9	99.9	100	100	100	100
86.7	76.6	52.1	34.0	9.0	4.7	1.0	0.1	90%	98%	98.5	99.3	99.8	99.9	100	100	100	100
86.4	76.0	51.4	33.3	8.8	4.6	0.9	0.1	90%	95%	96.4	98.2	99.4	99.7	99.9	100	100	100
85.7	75.0	50.0	32.1	8.3	4.3	0.9	0.1	90%	90%	93.1	96.4	98.8	99.4	99.9	99.9	100	100
84.2	72.7	47.1	29.6	7.5	3.9	0.8	0.1	90%	80%	87.1	93.1	97.6	98.8	99.8	99.9	100	100
81.3	68.4	41.9	25.5	6.2	3.2	0.6	0.1	90%	65%	79.4	88.5	95.9	98.0	99.6	99.8	100	100
76.9	62.5	35.7	20.8	4.8	2.5	0.5	0.0	90%	50%	73.0	84.4	94.2	97.2	99.4	99.7	99.9	100

Table II (contd)

Predictive value of a positive test result (%): estimated prevalence of infection*								Assay		Predictive value of a negative test result (%): estimated prevalence of infection*							
40%	25%	10%	5%	1%	0.5%	0.1%	0.01%	Diagnostic specificity	Diagnostic sensitivity	40%	25%	10%	5%	1%	0.5%	0.1%	0.01%
76.9	62.5	35.7	20.8	4.8	2.5	0.5	0.0	80%	100%	100	100	100	100	100	100	100	100
76.7	62.3	35.5	20.7	4.8	2.4	0.5	0.0	80%	99%	99.2	99.6	99.9	99.9	100	100	100	100
76.6	62.0	35.3	20.5	4.7	2.4	0.5	0.0	80%	98%	98.4	99.2	99.7	99.9	100	100	100	100
76.0	61.3	34.5	20.0	4.6	2.3	0.5	0.0	80%	95%	96.0	98.0	99.3	99.7	99.9	100	100	100
75.0	60.0	33.3	19.1	4.3	2.2	0.4	0.0	80%	90%	92.3	96.0	98.6	99.3	99.9	99.9	100	100
72.7	57.1	30.8	17.4	3.9	2.0	0.4	0.0	80%	80%	85.7	92.3	97.3	98.7	99.7	99.9	100	100
68.4	52.0	26.5	14.6	3.2	1.6	0.3	0.0	80%	65%	77.4	87.3	95.4	97.7	99.6	99.8	100	100
62.5	45.5	21.7	11.6	2.5	1.2	0.2	0.0	80%	50%	70.6	82.8	93.5	96.8	99.4	99.7	99.9	100
65.6	48.8	24.1	13.1	2.8	1.4	0.3	0.0	65%	100%	100	100	100	100	100	100	100	100
65.3	48.5	23.9	13.0	2.8	1.4	0.3	0.0	65%	99%	99.0	99.5	99.8	99.9	100	100	100	100
65.1	48.3	23.7	12.8	2.8	1.4	0.3	0.0	65%	98%	98.0	99.0	99.7	99.8	100	100	100	100
64.4	47.5	23.2	12.5	2.7	1.3	0.3	0.0	65%	95%	95.1	97.5	99.2	99.6	99.9	100	100	100
63.2	46.2	22.2	11.9	2.5	1.3	0.3	0.0	65%	90%	90.7	95.1	98.3	99.2	99.8	99.9	100	100
60.4	43.2	20.3	10.7	2.3	1.1	0.2	0.0	65%	80%	83.0	90.7	96.7	98.4	99.7	99.8	100	100
55.3	38.2	17.1	8.9	1.8	0.9	0.2	0.0	65%	65%	73.6	84.8	94.4	97.2	99.5	99.7	100	100
48.8	32.3	13.7	7.0	1.4	0.7	0.1	0.0	65%	50%	66.1	79.6	92.1	96.1	99.2	99.6	99.9	100
57.1	40.0	18.2	9.5	2.0	1.0	0.2	0.0	50%	100%	100	100	100	100	100	100	100	100
56.9	39.8	18.0	9.4	2.0	1.0	0.2	0.0	50%	99%	98.7	99.3	99.8	99.9	100	100	100	100
56.6	39.5	17.9	9.4	1.9	1.0	0.2	0.0	50%	98%	97.4	98.7	99.6	99.8	100	100	100	100
55.9	38.8	17.4	9.1	1.9	0.9	0.2	0.0	50%	95%	93.8	96.8	98.9	99.5	99.9	100	100	100
54.5	37.5	16.7	8.7	1.8	0.9	0.2	0.0	50%	90%	88.2	93.8	97.8	99.0	99.8	99.9	100	100
51.6	34.8	15.1	7.8	1.6	0.8	0.2	0.0	50%	80%	78.9	88.2	95.7	97.9	99.6	99.8	100	100
46.4	30.2	12.6	6.4	1.3	0.6	0.1	0.0	50%	65%	68.2	81.1	92.8	96.4	99.3	99.6	100	100
40.0	25.0	10.0	5.0	1.0	0.5	0.1	0.0	50%	50%	60.0	75.0	90.0	95.0	99.0	99.5	99.9	100

* An estimate of prevalence based on calculation of estimated prevalence from apparent prevalence (see Section entitled 'Estimating true prevalence from apparent prevalence'), or on an estimated prevalence in the population from which the samples were obtained

results will produce a PV– of 99.9% with a precipitous drop in PV+ to 9%. Given that decreases in prevalence affect primarily PV+, when prevalence falls it is desirable to move the cut-off to the right to increase DSp, for example to 99.9%; this will cause a commensurate drop in DSn to possibly 90%. However, this drop in DSn has a negligible affect on PV–; it will remain at 99.9% but will increase the PV+ from 9% to approximately 50% (11, 12).

Provision to clients of interpretation statements for test results

When test values are reported without providing estimates of the DSp and DSn of the assay, it is not possible to make informed predictions of infection status from test results. Hence, it is important that an interpretation statement accompanies test results. A small table indicating PV+ and PV– for a range of expected prevalences of infection in the target population is also useful, since clients are not likely to calculate predictive values from formulas. Without such information, clients will probably misclassify the infection status of animals; if that occurs frequently, the assay cannot be considered a fully validated assay.

Second part of the process: ensuring assay validity during routine use and enhancing assay validation criteria

Monitoring and maintenance of assay performance

The premise for this paper is that an assay is valid only to the extent that test results are valid. If the section entitled

'Establishing parameters and characterisation of assay performance' above (stages 1 to 3 in Fig. 1) is implemented, the conventional view of assay validation has been fulfilled. However, to assure valid results and to retain the designation of 'validated assay', constant monitoring, maintenance and enhancement of the assay are required. To extend the assay to disparate populations of animals, testing of reference animals representing those populations is required for updating estimates of DSn and DSp.

Precision and accuracy: the task of monitoring

Once the assay is in routine use, internal quality control is accomplished by consistently monitoring the assay using Levey-Jennings charts (Fig. 2) for assessment of repeatability and accuracy. Charts representing the last 30 runs will reveal trends or shifts in values of controls and standards. The lines representing ±1, ±2, and ±3 SDs from the mean can be used as decision criteria for inclusion or exclusion of one or several runs of the assay (16). The run is rejected if one control/standard exceeds ±3 SDs, or if two or more exceed ±2 SDs. Decision criteria may need to be customised for a given assay because of inherent differences between assays attributable to the host/pathogen system.

Proficiency testing

Reproducibility of test data between laboratories should be assessed at least twice each year. Membership of a consortium of laboratories that are interested in evaluating their output is valuable. In the near future, good laboratory practices including implementation of a total quality assurance programme, such as the International Organisation of Standards 9000 series (6, 7, 8, 9) and Guide 25 (10), will become essential for laboratories seeking to meet national and international accreditation requirements.

Proficiency testing is a form of external quality control for an assay. It is usually administered by a reference laboratory that distributes panels of samples, receives the results from the laboratories, analyses the data and reports the results back to the laboratories. If results from a laboratory remain within acceptable limits and show evidence of accuracy and reproducibility, the laboratory may be certified by government agencies or reference laboratories as an official laboratory for that assay. On the other hand, a laboratory that deviates significantly from expected values will not pass the proficiency test and will not be accredited. To maintain proof of the validity of the new assay, these steps are highly desirable. Panels of sera for proficiency testing should contain representation of the full range of analytes in the target population. If the panels only have sera with very high and very low values (with none near the cut-off point of the assay), the exercise will only give evidence of reproducibility at the extremes of analyte concentration, and will not clarify whether routine test results on the target population properly classify animals as to infection status.

Updating validation criteria

Due to the extensive set of variables that have an impact on the performance of serodiagnostic assays, it is useful to expand the number of reference sera when possible, due to the principle that error is reduced with increasing sample size. An expanded reference serum bank should be used to update estimates of DSn and DSp for the population targeted by the assay. Furthermore, when the assay is to be transferred to a different geographic region (e.g., from the northern to the southern hemisphere), it is essential to revalidate the assay by subjecting it to sera from populations of animals that reside under local conditions. Evaluating reference sera that represent those populations using stages 3 to 5 (Fig. 1) of the validation process will accomplish this requirement. It is the only way to assure that the assay is valid for populations that are of a different composition compared with the original population targeted by the assay.

Validation of new reagents or changes in protocol

When control samples are nearing depletion, it is essential to prepare and repeatedly test the replacement samples. The replacement samples should be included in at least 10 routine runs of the assay to ascertain their performance. When other reagents, such as antigen for capture of antibody, must be replaced, they should be produced or procured using the same protocols or criteria as used for the original reagents. They need to be assessed using sera from routine submissions in 5 to 10 parallel runs that include the current and the new reagent(s). Substituting an antigen produced by a different protocol will require full restandardisation and characterisation of diagnostic performance (stages 2 and 3 of assay validation: Fig. 1). Whenever possible, it is important to change only one reagent at a time to avoid the compound problem of evaluating more than one variable concurrently. These measures assure that the new reagents will not introduce excessive variability and assay validity should be maintained.

Validation of assays other than enzyme-linked immunosorbent assay

Although the example used has been an indirect ELISA test, the same principles apply to the validation of any diagnostic assay. It is extremely important not to stop after stage 2 of assay validation. That may result in a paper for the literature, but does not constitute a validated assay for diagnostic use. Although reagent and protocol workups during stages 1 to 3 are important, the selection of the reference populations is probably the most critical factor. It comes as no surprise when reviewing the literature to find a wide range of estimates for DSn and DSp for the same basic assay. Although part of the variation may be attributed to the reagents, more than likely

the variation in estimates of DSn and DSp is due to biased selection of sera upon which the test was 'validated'. Stages 4 and 5 in assay validation (Fig. 1) need more attention than they have been given previously. This is particularly true in the current atmosphere of international trade agreements and all the implications therewith regarding movement of animals and animal products.

Conclusions

With increasing trade and comprehensive trade agreements, importing countries need assurance that animals and animal products are free from certain disease agents. Testing of such animals must be performed with valid assays or no assurance of infection status can be established. Although laboratory accreditation is one mechanism for addressing this issue, there is no certainty that accredited laboratories are using validated assays. Fully licensed commercial assays may meet certain regulatory standards, but seasoned laboratory diagnosticians know that these assays are not always properly validated. Internal and external quality assurance programmes provide a mechanism for monitoring assays that may prove repeatable, reproducible, precise and even 'accurate.' But accuracy is a term that is relative to the 'standard of comparison' upon which the assay was based. If the standard is not valid, then the assay likewise is not valid. It is apparent that the first and foremost requirement for laboratory diagnosis of animal diseases or pathological conditions is a properly validated assay.

The ultimate goal of assay validation is to provide a test result that identifies animals as positive or negative, and by inference accurately predicts the infection status of animals with a predetermined degree of statistical certainty. Therefore, assay validation is a complex process that does not end with a time-limited series of experiments based on a few reference samples. The process also requires verification by application of the assay to a large number of reference animals that fully represent all variables in the population targeted by the assay. It also requires an interpretation of the data in a biologically and statistically relevant context. Only then can one gain assurance that the test result and the interpretation of that result correctly classify the infection status of an animal. This paper represents one perspective of assay validation procedures. Certainly there are valid points and counterpoints that need to be added to this perspective. May the discussion begin so that a consensus document on assay validation can be finalised.

Acknowledgements

The personnel of the OIE Reference Laboratory for ELISA and Molecular Techniques in Animal Disease Diagnosis at the Animal Production Unit, International Atomic Energy Agency (IAEA) Laboratories in Seibersdorf, Austria, and the affiliated Joint Food and Agriculture Organisation/IAEA Section for Animal Production and Health in Vienna are gratefully acknowledged for their many contributions to this paper and their endorsement of its content. Drs Peter Wright, Matthias Greiner and Susan Sutherland are also acknowledged for very constructive additions to the manuscript.

This paper is an expansion of the 'Principles of Validation of Diagnostic Assays for Infectious Diseases' published in the OIE *Manual of Standards for Diagnostic Tests and Vaccines* (17). The generalisations presented in that paper are detailed here to bring more clarity to the issues and essentials involved in assay validation.

■

References

1. Cembrowski G.S. & Sullivan A.M. (1992). – Quality control and statistics. *In* Clinical chemistry: principles, procedures, correlations (M.L. Bishop, J.L. Duben-Engelkirk & E.P. Fody, eds). Lippincott, Philadelphia, 63-101.

2. Crowther J.R. (1995). – ELISA theory and practice. *In* Methods in molecular biology. Humana Press, Inc., Totowa, New Jersey, 256 pp.

3. Greiner M., Franke C.R., Böhning D. & Schlattmann P. (1994). – Construction of an intrinsic cut-off value for the sero-epidemiological study of *Trypanosoma evansi* infections in a canine population in Brazil: a new approach towards unbiased estimation of prevalence. *Acta trop.*, **56**, 97-109.

4. 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. Meth.*, **185**, 123-132.

5. Hulley S.B. & Cummings S.R. (1988). – Designing clinical research. Williams & Wilkins, Baltimore, 139-150.

6. International Organisation for Standardisation (ISO) (1993). – Quality management and quality assurance standards. Part 4: Guide to dependability programme management. ISO 9004. ISO, Geneva, 21 pp.

7. International Organisation for Standardisation (ISO) (1994). – Quality management and quality assurance standards.

Part 1: Guidelines for selection and use. ISO 9001. ISO, Geneva, 18 pp.

8. International Organisation for Standardisation (ISO) (1997). – Quality management and quality assurance standards. Part 2: Generic guidelines for the application of ISO 9001, ISO 9002 and ISO 9003. ISO 9002, Geneva, 26 pp.

9. International Organisation for Standardisation (ISO) (1997). – Quality management and quality assurance standards. Part 3: Guidelines for the application of ISO 9001: 1994 to the development, supply, installation and maintenance of computer software. ISO 9003. ISO, Geneva, 32 pp.

10. International Organisation for Standardisation (ISO)/ International Electrotechnical Commission (IEC) (1990). – General requirements for the competence of calibration and testing laboratories. ISO/IEC Guide 25. ISO/IEC, Geneva, 7 pp.

11. Jacobson R.H. (1991). – How well do serodiagnostic tests predict the infection or disease status of cats? *J. Am. vet. Med. Assoc.*, **199**, 1343-1347.

12. Jacobson R.H. (1996). – Assessing the validity of serodiagnostic test results. *Semin. vet. Med. Surg. (Small Animal)*, **11**, 135-143.

13. Jacobson R.H., Shin S.J., Rossiter C.A. & Chang Y.-F. (1993). – Paratuberculosis: new diagnostic approaches for a difficult problem. *Aust. Microbiologist*, **14**, A53.

14. Kemeny D.M. & Challacombe S.J. (1988). – ELISA and other solid phase immunoassays – theoretical and practical aspects. John Wiley & Sons, New York, 367 pp.

15. McNeil B.J., Keeler E. & Adelstein S.J. (1975). – Primer on certain elements of medical decision making. *N. Eng. J. Med.*, **259**, 211-226.

16. MacWilliams P.S. & Thomas C.B. (1992). – Basic principles of laboratory medicine. *Semin. vet. Med. Surg. (Small Animal)*, **7**, 253-261.

17. Office International des Epizooties (OIE) (1996). – Principles of validation of diagnostic assays for infectious diseases. *In* Manual of standards for diagnostic tests and vaccines, 3rd Ed. OIE, Paris, 8-15.

18. Rogan W.J. & Gladen B. (1978). – Estimating prevalence from the results of screening test. *Am. J. Epidemiol.*, **107**, 71-76.

19. Smith R.D. (1991) – Clinical veterinary epidemiology. Butterworth-Heinemann, Stoneham, Maryland, 223 pp.

20. Tijssen P. (1985). – EIA data processing. *In* Practice and theory of enzyme immunoassays (R.H. Burdon & P.H. van Knipperberg, eds). Elsevier Science Publishing Company, New York, 385-421.

21. Wright P.F., Nilsson E., Van Rooij E.M.A., Lelenta M. & Jeggo M.H. (1993). – Standardisation and validation of enzyme-linked immunosorbent assay techniques for the detection of antibody in infectious disease diagnosis. *In* Biotechnology applied to the diagnosis of animal diseases. *Rev. sci. tech. Off. int. Epiz.*, **12** (2), 435-450.

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