Rev. sci. tech. Off. int. Epiz., 2006, **25** (3), 913-935

Aspects of kit validation for tests used for the diagnosis and surveillance of livestock diseases: producer and end-user responsibilities

J.R. Crowther, H. Unger & G.J. Viljoen

Animal Production and Health Section, Joint Food and Agriculture Organization (FAO)/International Atomic Energy Agency (IAEA) Division of Nuclear Techniques in Food and Agriculture, IAEA, Wagramer Strasse 5, A1400 Vienna, Austria

Submitted for publication: 16 January 2006
Accepted for publication: 22 March 2006

Summary
The Joint Food and Agriculture Organization/International Atomic Energy Agency (IAEA) Division of Nuclear Techniques in Food and Agriculture, based at the IAEA in Vienna, Austria, has extensive experience in helping to develop and validate assays and has provided strong support in developing World Organisation for Animal Health (OIE) norms. This paper will focus on enzyme-linked immunosorbent assay and polymerase chain reaction as the major technologies exploited in diagnosis and surveillance. Problems involving the terminology and factors in kit production, supply and validation are examined, in particular emphasising the importance of robustness and ruggedness of tests. The authors discuss the responsibilities of the various stakeholders (producers, distributors, users, and national/international organisations) in achieving quality controlled data to solve diagnostic and surveillance problems. The roles of internal quality control (internal proficiency testing) and external quality assurance (external proficiency testing) as well as aids to solving problems with kits are examined.

Keywords
Diagnosis – Diagnostic sensitivity – Diagnostic specificity – Enzyme-linked immunosorbent assay – Guideline – Kit – Polymerase chain reaction – Robustness – Ruggedness – Surveillance – Validation – World Organisation for Animal Health (OIE).

Introduction

The latest World Organisation for Animal Health (OIE) guidelines for validation of tests have been developed in conjunction with the Food and Agriculture Organization (FAO), the International Atomic Energy Agency (IAEA) and the scientific community. The guidelines define stages of validation applied to a particular test's specific fitness for purpose and form the basis of the OIE's peer-reviewed test registration and certification process. Whilst recognising that validation is a continuous process, the OIE defines four stages (one to four) that gradually expand data to increase confidence that a test is valid for a wider range of countries and laboratories (see 'Certification of diagnostic assays' on the OIE homepage – www.oie.int). There is still, however, debate concerning the definition of certain terms, in particular, 'ruggedness' and 'robustness'. Neither is popular for discussion since areas have to be considered that are difficult to quantify in terms of a test performance. Most opinions seem to define robustness as a measure of the resistance of a test to being affected by physical factors and ruggedness as a measure of how a test performs under a wide variety of operational conditions. Ultimately, tests for use in diagnosis and surveillance have to be judged by their diagnostic performance, which is measured in terms of diagnostic sensitivity (DSn) and diagnostic specificity (DSp). Here DSn refers to the proportion of known infected reference animals testing positive [true positive/(true postive + false negative)], while DSp refers to

the proportion of uninfected reference animals that test negative [true negative/(true negative + false positive)]. The use of samples from representative populations is a major difficulty in validation. It is worth contrasting analytical sensitivity (ASn) and analytical specificity (ASp) measurements in validation, which are determined with more defined reference materials (e.g. from experimentally derived samples). Here ASn refers to the ability of a test to measure the amount of analyte, such as antibody or antigen(s) in a reference material(s) and ASp refers to the ability of a test to measure that analyte specifically in the presence of similar defined substances. Test DSn and DSp are factors of the performance which are derived from testing a specific population with a known history and infection status. It is important to use reference materials (standards) to keep track of a test by examining the variations in ASn and ASp, this confirms whether or not the basic components of a test system are functioning to a required level of performance, which is particularly important in internal quality control (IQC). In order for kits to be reliable they must be resistant to physical forces and provide reagents and protocols that enable consistency to be maintained, i.e. they must be stable.

Stability of kits/tests

The terms ruggedness and robustness deal with test stability from two angles:

– physical factors affecting test performance (robustness)

– factors affecting repeatability in a laboratory and reproducibility between laboratories (ruggedness).

Physical factors possibly impairing the biological function of the reagents in a test

The principal physical factors that influence test robustness are storage conditions during transportation to the end-user (and time of storage before receipt), storage conditions in the user's laboratory, and repeated use of the test. These factors should be grouped under the term 'robustness'. So, a robust test is one where the reagents are resistant to a wide variety of physical factors during transportation and storage. A non-robust test would contain one or more elements that were easily affected by the same conditions. An example would be an enzyme conjugate whose activity was known to be drastically affected by relatively small increases in temperature above 4°C. Such a reagent might be shipped in ice, but there is always the possibility of delays in transportation, which will mean that the temperature cannot be maintained. This is even more marked when reagents have to be sent in dry ice. Although the sender may have supplied specific instructions for transportation, if there is a delay, the test is subject to deterioration. There may be several reagents that are sensitive to rises in temperature and each element affected increases the chance that the test will not perform as expected (i.e. as it would have done when it left the supplier) due to its 'non ruggedness'.

Assessing the stability of all reagents is not easy since all possible conditions cannot be predicted or tested. The sustained supply and successful use of a kit is a complex problem. The developmental stages are often part of the validation process *per se* and it is sometimes difficult to assess at what stage a kit is at in the overall assessment. Figure 1 shows some basic features of kit supply. Alternative pathways for supply are shown either as direct links between producer and user through transportation; or through links between producers and users via a distributor. The kit has to be 'fit for use' when it leaves either the producer (with full quality control [QC]) or the distributor (where QC testing may not be done). At this stage, the kit meets the 'fit for use' criteria, implying that when the kit leaves the producer's or distributor's premises it is formulated to achieve the defined performance. The possibility of a kit being affected by physical factors is increased with each round of transportation.

Conditions during kit transportation are not as controllable as they are in the laboratory and therefore not as quantifiable. Factors such as temperature changes, shaking, leakage, dehydration, ultra violet light, etc., that might affect the performance of a kit, are not easy to reproduce in supplier's laboratories and their effects are not easy to measure. Moreover, features of test reagents that are inclined to be subject to the effects of temperature and storage time under experimental conditions are likely to be even less robust in practice. Although precautions to avoid such problems are advised, they are not always followed, or there is no awareness, or acknowledgement, of adverse conditions. Table I lists possible effects on biological reagents during transportation. Anything that affects a kit in this phase can be considered as influencing the robustness of the kit.

In addition, once received by the user, the kit can be affected by laboratory variations in storage and handling and this should also be considered under robustness. It might be useful to think about defining robustness in terms of 'transport robust' and 'laboratory robust'. The stability of a test can be affected by conditions of storage and handling in the laboratory and during transport, but in the laboratory, because conditions can be more easily controlled and identified, it is possible to determine (using IQC to evaluate laboratory proficiency) the exact conditions which led to the sudden failure of a test or to a slow deterioration in performance. This will be discussed in more detail later.

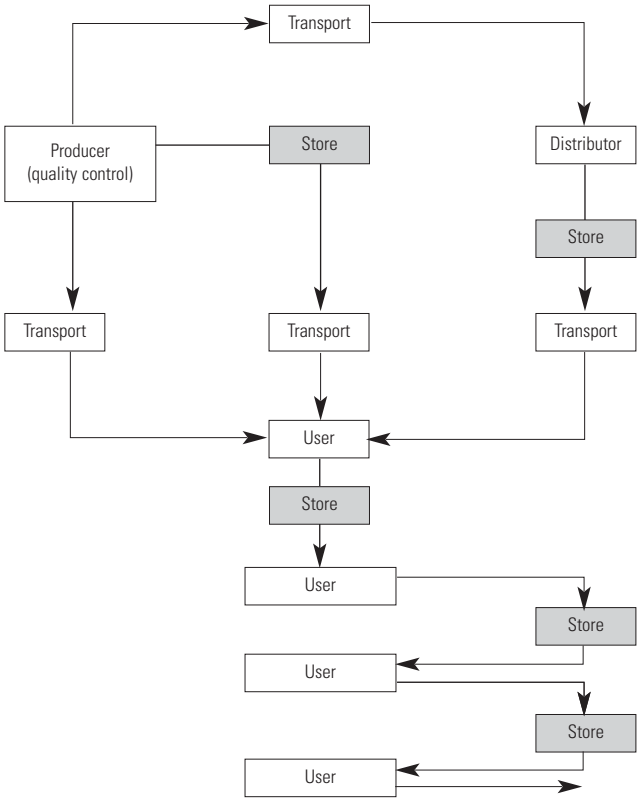


Fig. 1
Kit supply: the reagents in a test kit must be 'robust' enough to withstand transport, storage, and repeated use
The possibilities of reagents deteriorating are increased by the number of stages in transportation and storage

Test repeatability and reproducibility in terms of the widespread use of the test in conditions which are variable and less than ideal

Where a test functions as expected on receipt it can be described as having a high degree of ruggedness, i.e. it has features that resist change through the many possible variations in technique. The ultimate test is absolutely foolproof, can be done by totally untrained individuals, gives the same results for controls every day, every month and every year (highly precise), and can accommodate large variations in experimental handling and technique. In other words, ruggedness is a measure of resistance to user variables. It is not easy to monitor performance and cover the ability of staff to follow given instructions. Examination of the OIE validation criteria shows that repeatability and reproducibility (i.e. the ruggedness of the tests) are measured at Stages 2 and 3 of the process, respectively. It can be predicted that the more complex a test is and the more variables involved, then the less rugged it is likely to be.

Samples used for diagnosis

The performance of any system is determined with reference to a defined set of reagents (e.g. a kit). The function of a test is to do a certain job, and a test's ability to perform this job is described by the OIE as 'fitness for purpose'. Users should be provided with reagent kits that include the necessary controls to allow measurement of an analyte so that a conclusion can be drawn as to whether or not the test is fit for its stated purpose. Attention should be paid in terms of the ruggedness or robustness of any system as to the influence of storing and processing on the test sample; this can have a profound effect on results. If the results of a test are easily affected by the conditions under which the samples are taken then such a test cannot be described as 'rugged'. For instance, if a sample has to be taken and immediately placed in liquid nitrogen and stored for exactly 11 h and 54 s. before being tested, plainly this is a ridiculous and totally untenable method. Such a test system is not rugged since the conditions cannot be met. The taking of samples, their analysis by a test, and the effect of different procedures should all be included in the validation data. For polymerase chain reaction (PCR), the sampling and processing (nucleic acid extraction) is extremely important to DSn determinations and slight alterations in technique or equipment used can greatly alter the diagnostic potential of a test.

Kits and reagent sets

Defining a kit is useful since reagents and protocols proclaiming to be kits can be very different and kit formulation can drastically affect their performance in terms of ruggedness and robustness and also in solving problems. There is a large distinction between the concepts of a kit for enzyme-linked immunosorbent assay (ELISA) and a kit for PCR. Kits for PCR will be discussed in a later section highlighting the differences.

The definition of what comprises a kit rests on the following considerations:

– test validation
– the perceived objective of the kit
– the 'market' or end-users who are to exploit the kit
– factors involved in sustainability.

Producing a kit is complex and involves issues of technical performance, supply, profit motives and continuity. Kits have to be accepted by international bodies if they are to fulfil their ultimate role of standardising a given approach to evaluating a given situation and allowing harmonisation with other tests measuring the same or similar factors.

Table I
Factors affecting biological reagents in kits on transportation

Adverse factors	Effects	Comments
Low temperature		
Freezing	Concentration	Affects activity of reagents through errors in concentration
	Phase separation	
Freeze thawing	Denaturation proteins	Affects biological activity (both quantitative and qualitative)
	Coagulation	Clumping and aggregation can affect activities and increase dilution errors
	Inactivation enzymes	Reduces or eliminates biological activity
	Differential concentration of proteins	Affects concentration of reagents
High temperature	Denaturation proteins	Reduces or destroys biological activity. Can selectively alter quality of reagents
	Inactivation complement and other factors	Affects results on variably heated samples
	Dehydration	Concentration of reagents. Could denature proteins
	Evaporation	Loss of fluid, so test volumes too low
	Contaminant growth	Proteolysis, bacterial and fungal activity
	Hydration with humid conditions	Dilution
		Destruction of freeze-dried reagents
Shaking	Frothing	Phase separation
	Vigorous shaking	Shearing effects on proteins (inactivation)
Ultra-violet inactivation, sunlight	Inactivation	Biological activity altered or destroyed
Leakage	Loss of necessary volume	Reduces number of samples for test
	Contamination	Can lead to microbial degradation of reagents
Breakage	Loss of reagents	Reduces test numbers
	Contamination	Chemical or microbial breakdown/inactivation
	Mixing	Cross-contamination of activities
Re-hydration	Freeze-dried reagents affected by moisture	Can destroy or alter biological activity directly or after microbial contamination

The definition of an ultimate kit may be formulated by examining kits that are already in use and such a definition can be used to help design better kits. Having said this, there is no perfect kit that deals with biological systems. The gathering of information from kits and the modification of reagents/conditions/protocols are necessary to account for the many variables which cannot be assessed at a single time point. Validation also involves changes in biological systems, such as alteration in the antigenicity of the agents examined, which necessitates action.

A perfect kit

a) A kit should contain everything needed to allow testing, including software to facilitate the training of laboratory personnel, the storage and processing of samples, and the analysis and reporting of data

b) the reagents should be absolutely stable under a wide range of temperature conditions

c) the manual describing the use of the kit should be 'foolproof'

d) the kit should be validated 'in the field' as well as in research laboratories

e) all containers for reagents should be leak-proof

f) IQC samples should be included

g) external quality assessment should be included in the kit 'package'

h) data on the relationship of kit results to those from other assays should be included

i) attention should be given to ensuring that all equipment used in association with the kit is calibrated (spectrophotometers, pipettes)

j) training courses in the use of the kit should be organised

k) information exchange should be set up to allow rapid 'on-line' help and evaluation of results where there are perceived problems

l) the internationally agreed standards for the supply and control of kits should be maintained.

The establishment of these perfect conditions would greatly reduce problems in the ruggedness and robustness of tests. The FAO/IAEA experience is that producers are very variable in their attention to the above criteria. The main criticisms are shown in Table II, where points 1 to 12 refer to the perfect kit criteria above. Not surprisingly,

Table II
Problem areas regarding available enzyme-linked immunosorbent assay (ELISA) kits

Perfect kit criteria	Performance of available kits
1. Completeness	Lack of micropipette tips and of the high-quality water needed for initial reagent dilutions are the main problems for ELISA
2. Stable reagents	Some problems
3. Foolproof manual	Variable quality. Generally incomplete background
4. Field-validated	Can be poor (hence the need for World Organisation for Animal Health guidelines)
5. Leak-proof	Good
6. IQC samples and methods	Some kits have necessary controls missing in their regime
7. EQA responsibilities	Poor
8. Relationship data to other tests	Not usual
9. Calibration equipment	Left to the user
10. Training courses	No
11. Rapid reporting problems	Not many systems in place
12. Standards	Poor

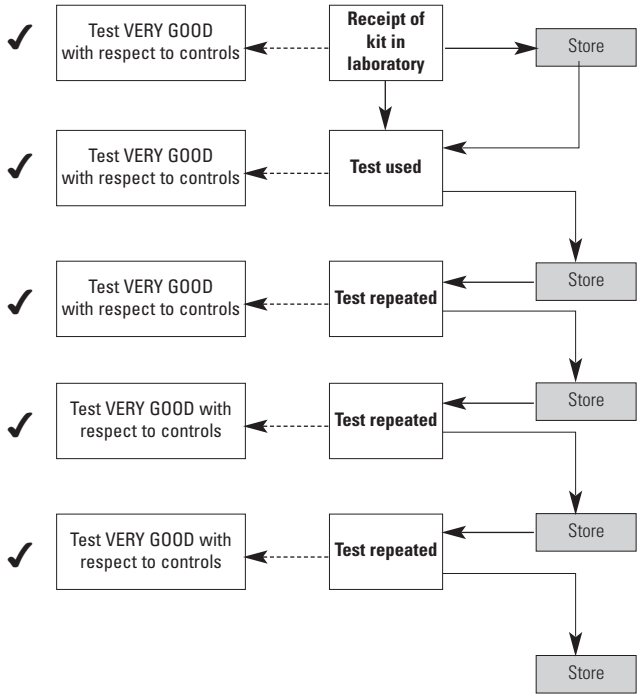
EQA: external quality assurance
IQC: internal quality control

most of the features are expressed in the test performance validation criteria required by the OIE.

Producer/end-user responsibilities

Features of robustness and ruggedness are illustrated in Figures 1, 2, 3 and 4. Figure 1 shows the various pathways for transporting kits. Figure 2 illustrates a perfect scenario in which a kit 'survives' transportation and delivery as judged by the successful confirmation that the test works. After storing and re-using the test there are no problems with the controls provided. In this scenario there are obviously no problems with transport robustness or laboratory robustness or ruggedness. The IQC monitoring of test performance is vital in assessing performance.

Figure 3 illustrates a scenario where the kit is received but does not work, presumably because some of the elements are faulty. If the assumption is made that the quality control of the supplier was good and it left as 'fit for use', then it might be concluded that transportation may have affected performance (non robust). It is worthwhile for the user to repeat the test since it may have been an operator error which led to the poor performance. The major causes of problems with a new kit (or even technology) are operator errors in dilution and manipulation and non-adherence to protocols. The kit performance is related to the controls provided. On confirmation that something is wrong on receipt, there are choices for the user. He/she may contact the producer and report the finding to ask for help or a new kit. The responsibility to gather, respond and update data on their test is that of the producer and this is included in Stage 3 of the OIE validation guidelines. The



✓ = good test

Fig. 2
Test ruggedness/robustness on receipt and after use – Scenario A: no problems in robustness and ruggedness
The test is monitored with respect to controls and works every time. This is the role of internal quality control

responsibility of the user is to provide the producer with data that is quality-controlled and to follow the given protocols closely.

The user can also examine which of the test reagent(s) is causing the problem(s) through trouble-shooting. This requires a good level of understanding of the principles of

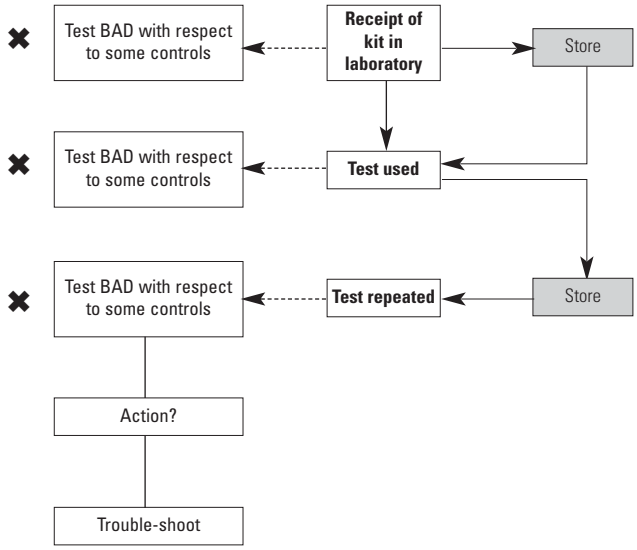


Fig. 3
Test ruggedness/robustness on receipt and after use – Scenario B: possible ruggedness problem
A problem is identified on receipt of kit. Re-testing kit gives same result. Actions can be taken to investigate why the kit has failed with respect to controls, including examining the controls themselves, since not all may have failed. Trouble-shooting may allow test reagents to be rescued, but this depends on the 'skills' of end-users, and the ideal conditions recommended by the producer would not be met

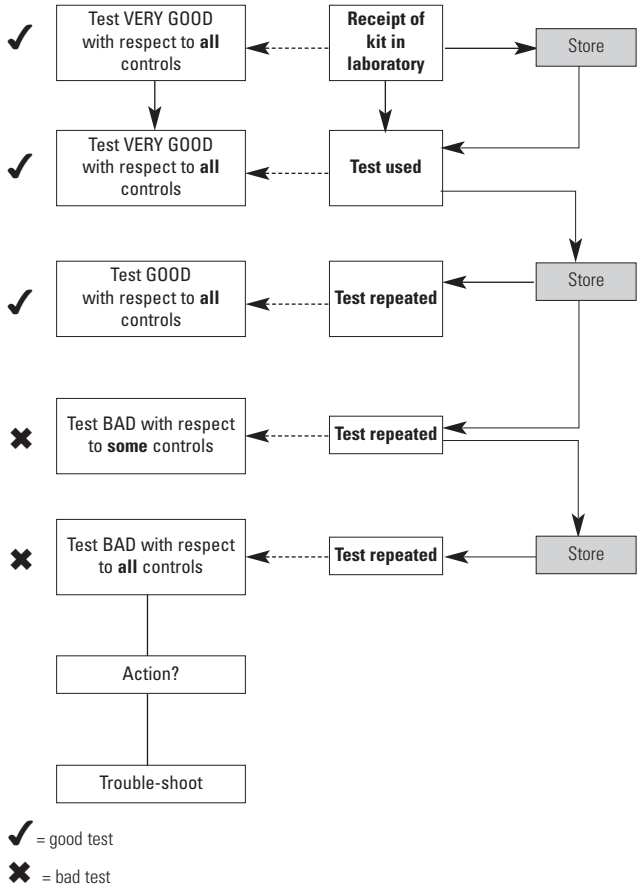


Fig. 4
Test ruggedness/robustness on receipt and after use – Scenario C: gradual changes in performance
This is monitored through internal quality control with respect to given controls. The test performs well at the beginning but either gradually changes with respect to controls or is found to be failing at a specific time. Trouble-shooting may identify which reagent is failing and whether this can be compensated

the tests being used and the ability to think clearly and perform accurate experiments. An example of trouble-shooting when using an indirect ELISA is given later to illustrate that the process can be quick and not too expensive in terms of reagents. The way forward will depend on the results of the trouble-shooting. If a single reagent is 'faulty', e.g. has lost activity, then this might be replaced by the supplier or its concentration adjusted. If controls (one or more) are affected it is difficult to replace these to achieve the producer-defined characteristics of ASn and ASp.

Figure 4 shows a third scenario, in which a kit works initially but, at a later time, fails. The constant monitoring of a kit's performance with respect to controls is again emphasised, e.g. by use of charting methods to continuously plot control data, so that trends in data can be seen. Failures may be progressive (through observation of a trend), or more sudden. Examination of which reagents are affecting a test performance is vital to ascertain what steps can be taken to bring the test back into use. As with the considerations above, single reagents can be obtained from the producer, e.g. controls. The use of in-house controls, e.g. a control positive serum, if included in tests as extra controls, can also be useful to measure the quality of the controls provided, e.g. the in-house control may retain its expected values whereas the controls provided in the kit may have reduced values over time. Such aspects come under the heading IQC. Robustness factors are more likely to be seen on delivery or soon after use. Ruggedness factors are linked strongly to good laboratory practice (GLP) and the precision at which the kit functions under a wide range of variables.

Relevance of ruggedness and robustness to the OIE validation pathway

The producer is responsible for maintaining a test's robustness throughout transportation. Physical factors

affecting robustness are also present when the kit is received, but they are then under the control and responsibility of the laboratory. Assessing robustness means assessing the stability of reagents over a range of conditions. Where a kit has been examined in many laboratories worldwide and attention has been paid to stabilising reagents for travel and storage in a laboratory, then data will be available to measure the success of a test under those conditions. During development the producer may have to adjust conditions to increase the stability of reagents and the eventual kit formulation can then be deemed robust. This is a major part of the validation process and can be costly in time and resources.

It is possible to obtain robustness data in the shorter term (predictive) by accelerating possibly adverse conditions, but this is difficult, and in most cases not necessary, because usually enough data will already be available on the components of the kit for robustness to be expected (even if the robustness of the complete kit itself is not initially tested). So, although no assumptions can be made until kits are sent out, the prediction that it is stable can be made.

The OIE guidelines state that the ruggedness and robustness of a test should be examined, but exactly what quantifiable data is needed to measure this is not defined, which will lead to confusion and more subjective approaches. Data might be best taken from several sources (examples are given in Table III).

Enzyme-linked immunosorbent assay: trouble-shooting

When a set of reagents in the form of a kit is sent to the user laboratory, any of the situations in Figures 2, 3 and 4 could arise. In all cases, examination as to whether the reagents are behaving as expected by the producer is performed with reference to given controls (at least initially) and through strict adherence to the given protocols. The different stages in this examination process (as shown in Figure 5) are described below.

Stage 1

The kit is opened and the reagents carefully made up and/or used as instructed. The test is performed exactly as described and the control optical density (OD) values and processed values calculated. One of two things will happen at this point:

Table III
Quantifiable factors of robustness and ruggedness

Factor	Data
Experience with sending out kits	Time over which kits supplied (supplied before the expiry date)
	Number of kits sent and number of batches sent
	Mode of transportation
	Adverse factors reported, e.g. delays at airports
Experience from a single laboratory	Number of labs where IQC was good on receipt
	Number of labs where IQC was not good on receipt
	Number of labs where IQC data alters from good to bad
Cumulative data from one laboratory	Unprocessed data collected and analysed
	IQC data analysed
	Repeatability data
Cumulative data from many laboratories	Unprocessed data collected and collated
	External quality assurance data from supplier
	External quality assurance data from users
	Statistical analysis of data from number of laboratories
	Repeatability data compared (statistic)
	Reproducibility data shown (statistic)
Accelerated shelf life determination on one or more reagents	Data
Problems reported	Number and extent
Problems solved	Number and extent
Changes necessary in protocols to establish robustness and ruggedness	Show data on solving problems to achieve higher robustness and ruggedness
Sampling and samples	Data on variations in samples and extent to which they affect tests

IQC: internal quality control

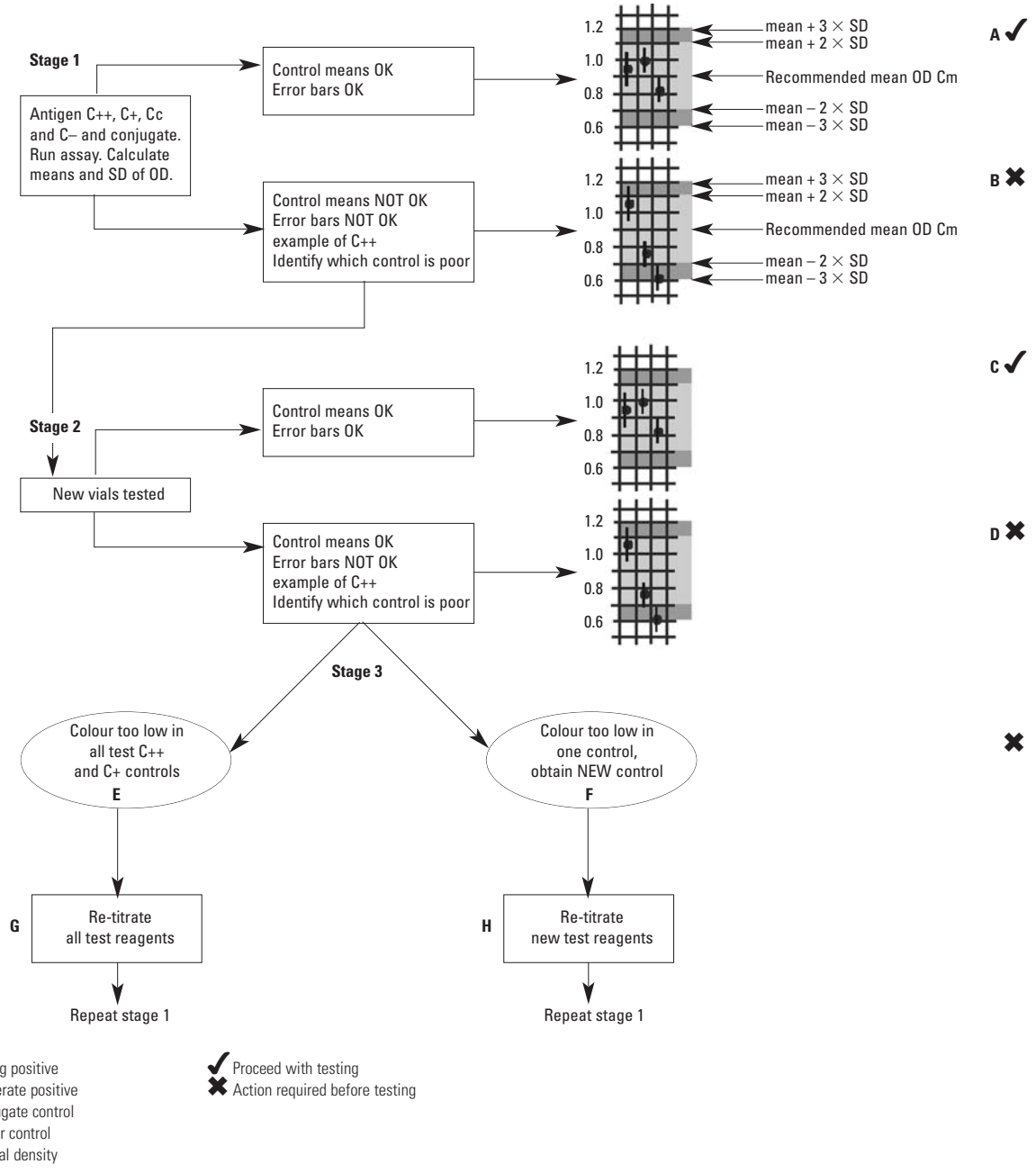


Fig. 5

Stages in examining an enzyme-linked immunosorbent assay kit and consequences of different results

The results for the various controls are plotted. The limits recommended for the values are shown, as well as the mean and standard deviation (SD) from the mean of the results for controls. Such control charts are ideal for continuous monitoring (internal quality control)

Scenario A: the test works exactly as 'expected' with means and error bars within limits. There is no reason to postpone testing of test samples (situation in Fig. 2)

Scenario B: one or more of the controls do not meet expectations either with respect to the mean values or high variation (situation in Fig. 3). If this situation occurs the user laboratory should proceed with Stage 2 of the examination.

Stage 2

In the case of experiencing a problem with one or more of the controls, the test should be repeated with a set of new vials of the reagents. When re-tested the assay may work as originally expected. If however, problems re-occur (as shown in scenario D, Fig. 5) then Stage 3 of the examination should be performed.

Stage 3

This depends on the extent of the problem(s) and whether the problem(s) can be associated with one particular control. For example, in scenario E (Fig. 5) both the strong positive (C++) and the moderate positive controls (C+) provided could have a low OD. In this case all the reagents (antigen, controls, and antibody conjugate) should be re-examined. A plate design which can help trouble-shoot all reagents in a single plate is shown in Figure 6. The colours in any control could also be strong. Note that situation E could be observed in Stage 1 (as in B for both controls). In this case, Stage 2 should still be performed, since there may well have been a common problem. In situation F (Fig. 5) there is colour that is out of limits in only one control. Here, it is likely that, after obtaining the same result for two consecutive runs with separate vials of control, there is a fault in the control common to all vials. In this case, a new batch of that control, or a new kit, should be obtained.

Evaluation of reagents

A typical system for an indirect ELISA involves antigen, control antibodies and a conjugate to detect bound antibodies. The kit instructions usually fix the dilutions of each of these. A simple method of testing the dilution range of each reagent will help indicate whether there is a major problem with the activities of any of the reagents.

Figure 6 shows a microtitre plate design to examine the three parameters of antigen, control serum and conjugate using higher concentrations of reagents. The steps in the examination process are:

a) a plate is coated with antigen as shown. This is diluted at 4 ×, 2 × and 1 × the recommended concentration

b) plate is incubated and washed

c) the control sera are added as shown, starting with 8 × the recommended dilution and diluted twofold in blocking buffer

d) the plate is incubated and washed

e) the conjugate is added as shown in two blocks representing 4 × and 2 × the recommended dilution

f) after incubation, washing, and addition of substrate, the reaction is stopped at the recommended time and the plate is read.

The results of the test should indicate whether any of the reagents are not working. This does not account for the substrate/chromophore failure where there would be usually no colour developing over the entire plate even on extended incubation. Figure 7 shows possible results and conclusions to illustrate the usefulness of this system for an end-user faced with a 'failed' test.

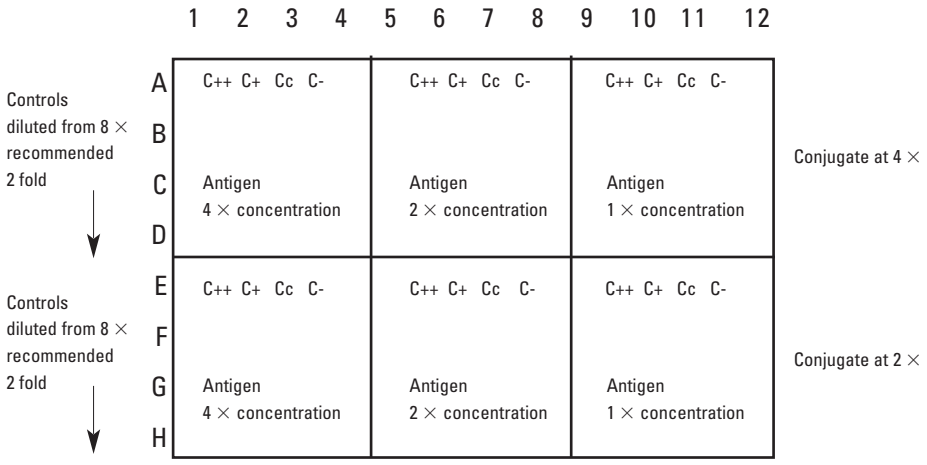
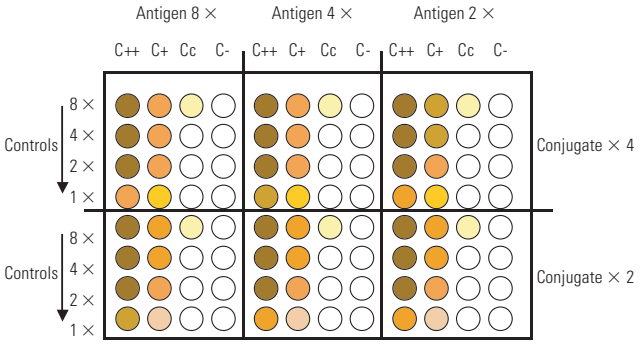


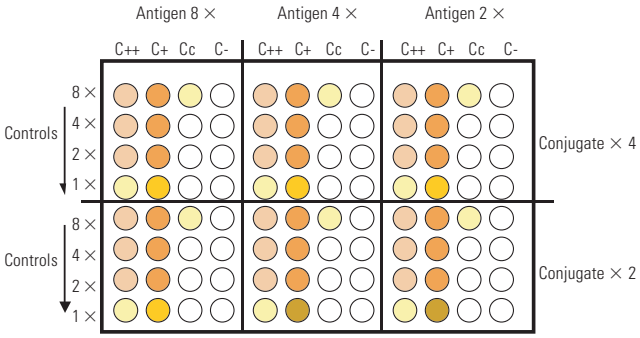
Fig. 6
Plate design assessing reagents for indirect enzyme-linked immunosorbent assay
Antigen is coated at 4 × recommended concentration and twofold dilutions in respective areas of plate. After incubation and washing, control sera are added diluted 8 times and twofold, in respective parts of the plate. In this way, a mini-chess board titration of antigen and control sera is obtained. After incubation and washing, the conjugate is added 2 × and 1 × the recommended working dilution. After incubation and washing, the chromophore/substrate is added as recommended. On development the plate is stopped (if this is instructed). The patterns of reactivity should allow identification of single reagents where there is a possible concentration or qualitative problem. The assessment of the chromophore/substrate is not made here, but can be checked as a separate issue

a) no problems



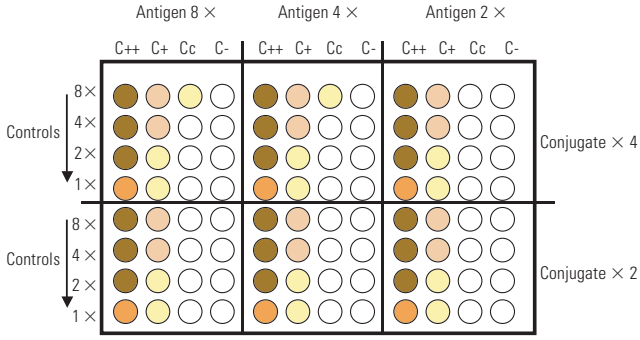
Both controls show strong development of colour reflecting increased concentrations of reagents

b) problems with the strong positive control

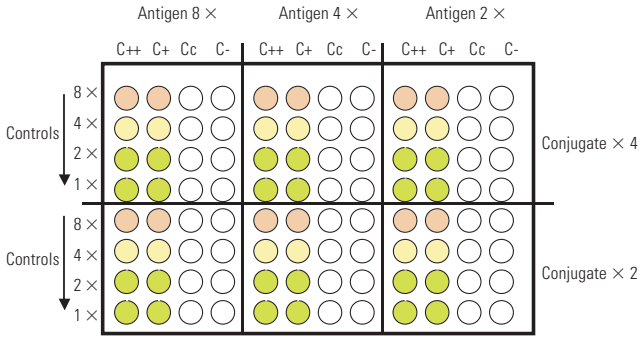


The optical density for the C++ is too low. The C+ results are as expected

c) problems with the moderate positive control

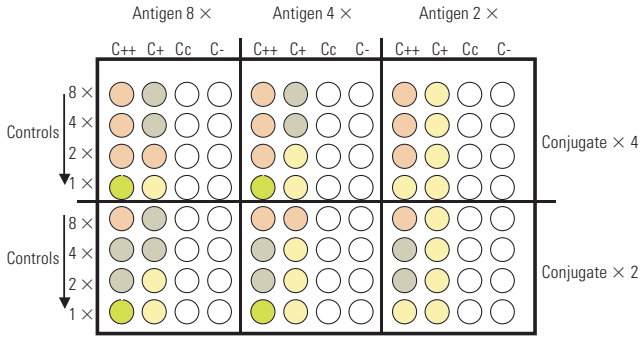


d) conjugate problems



There is little signal even for high concentrations of conjugate for both controls

e) antigen problems



This can be confused with the conjugate problems but generally there would be more colour and possibly a much reduced maximum plateau height for colour even in excess controls, positive sera and antigen

Fig. 7
Possible results and conclusions of reagent assessment

Use of laboratory sera

Where a laboratory has sera that have been proved positive in other tests then there is an opportunity to use this in the ELISA where problems with controls occur. These may be used, for example, where both controls are low, to ascertain whether the expected colour of the system can be achieved. Internal controls can also be included in continuous IQC.

Charting methods

The IQC results are obtained continuously and should highlight problems as they arise. It is imperative that operators process and examine results constantly and take the necessary actions as dictated by results. The use of charts to plot IQC data has already been indicated. A full treatment of charting methods can be seen on the IAEA web pages (indirect ELISA at http://www-naweb.iaea.org/nafa/aph/public/iqc-charts-elisa.pdf; competition ELISA at http://www-naweb.iaea.org/nafa/aph/public/iqc-charts-elisa-c1.pdf).

The objectives of charting data are:

– to keep a constant record of all data

– to monitor the assay from plate to plate in any one day's testing

– to monitor the tests made from day to day, week to week, year to year

– to allow rapid identification of unacceptable results

– to allow recognition of reagent problems

– to identify trends in results (increasingly poor performance)

– to identify when a new set of kit reagents is necessary

– to allow identification of differences in operators of the assay

– to fulfil various criteria for GLP

– to fulfil necessary requirements for external recognition that tests are being performed at an acceptable level (increasingly important where results are used for international trading purposes).

The documents on the IAEA website describe methods that allow test operators as individuals to monitor the performance of their test. Where there is more than one operator the method produces a unification of approach to allow results to be controlled and discrepancies to be identified. The documents are also a way of providing 'open' results that can be viewed by anyone, including outside scientists interested in evaluating the status of a laboratory involved in providing results on which management decisions concerning disease control are made. The methods described are not an in-depth statistical analysis of the data, but they attempt to visually assess results as descriptive statistics, to increase awareness of operators as to what they are performing on a daily, monthly and yearly cycle of work. The most important feature of testing in a laboratory is that operators have a very good understanding of the principles of the test they are performing and that they fully understand the nature of their results and the need to process data. There is no substitute for this understanding, but the charting method recommended is an aid to simplifying the process of test performance.

The charting methods have only two requirements:

– that the mean and standard deviation (SD) from the mean of the control samples' OD readings, and the mean and SD of the processed data (e.g. percentage positivity for indirect tests or percentage inhibition [for competitive tests]) are calculated with reference to the relevant given control values

– that the values obtained are plotted on two kinds of chart: the daily detailed data charts (DDD) and the summary data charts (SDC).

Thus, there is a systematic approach to data management that should impose a level of control on all laboratories using the same kit. Internal quality control data are an integral component of the external quality assurance programme (EQAP), since the results from any laboratory can be examined and correlated with results obtained after an EQA exercise, whereby the same limited number of samples are assessed at given times by all laboratories involved in a network of sero-monitoring or sero-surveillance.

The word 'transparency' is meant to indicate that results and processed data are available to all for comment. Ultimately, a set of well-presented 'good' results is a credit to a laboratory and engenders good team spirit and sustains interest in what can be a mundane task (continuous testing). Early indication of problems which can usually be easily solved (obtain new kit or new conjugate, change water, re-train operator, etc.), is a credit to the system used to assess performance and ultimately saves a great deal of time and resources. In the past, assays using kits have been run poorly in laboratories for a long time and the results have been used in disease management. The late identification of bad assays destroys faith in the kit, reduces managers' faith in the competence of laboratory staff and worst of all produces bad management decisions. Three examples of charting results are shown in Figures 8, 9 and 10.

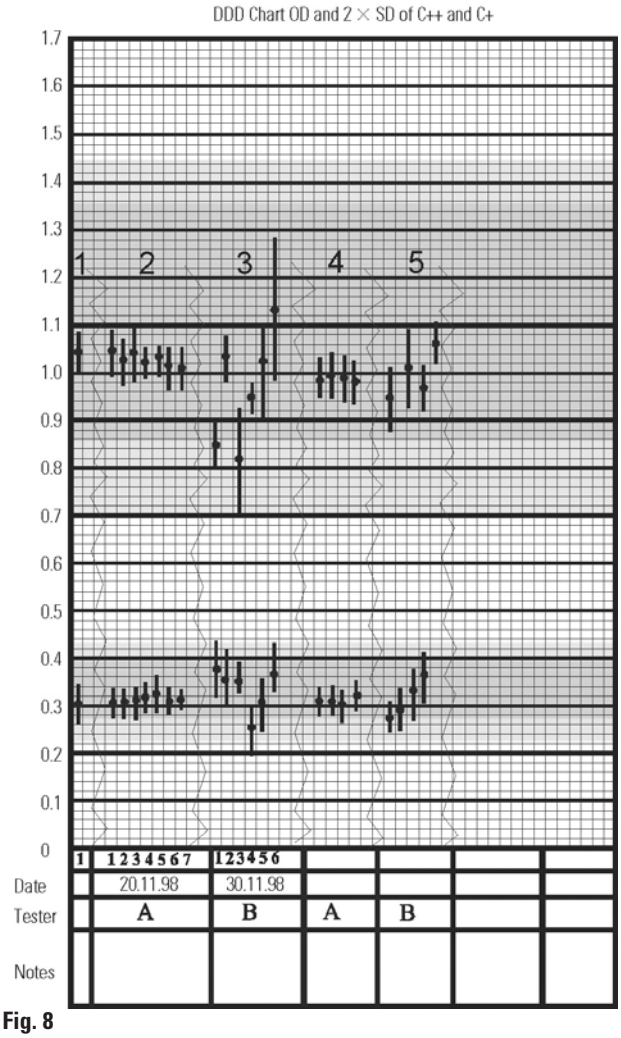
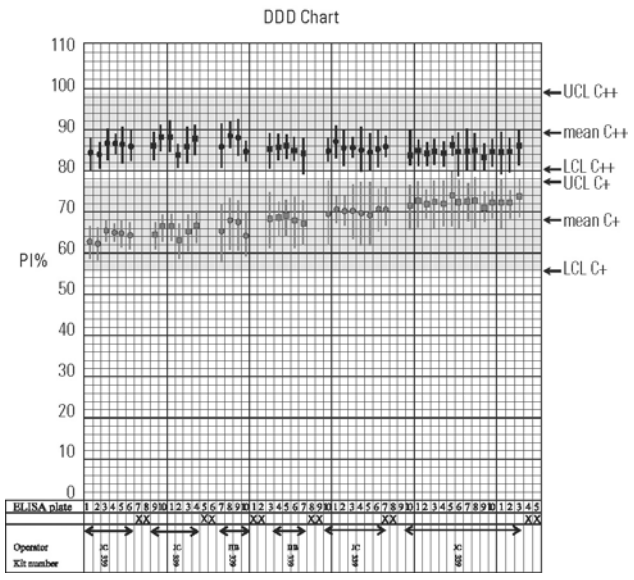


Fig. 8
Daily detailed data (DDD) chart showing plots from five tests of an indirect enzyme-linked immunosorbent assay of control sera activity for strong control (C++) and weaker control (C+) sera

Each plate in a test has data plotted. The tester and date are included on the chart. Test 1 had 1 plate; test 2 had 7 plates; test 3 had 6 plates; test 4 had 4 plates and test 5 had 4 plates. The mean value of the optical density (OD) for the controls is shown for C++ and C+ for each plate (plotted as black points). The allowable variation from the expected OD of the controls (as stated by the kit supplier) is shown in grey for both C++ and C+. The bars show 2 standard deviations from the mean for each point (variation). NB, in test 3 there are problems with means that are out of the allowable range for some plate controls as well as higher variation with respect to the length of error bars.



UCL: upper control limit
LCL: lower control limit
PI: percentage inhibition

Fig. 9
Daily detailed data (DDD) chart showing plots from six tests of a competitive enzyme-linked immunosorbent assay of control sera activity for strong control (C++) and weaker control (C+) sera

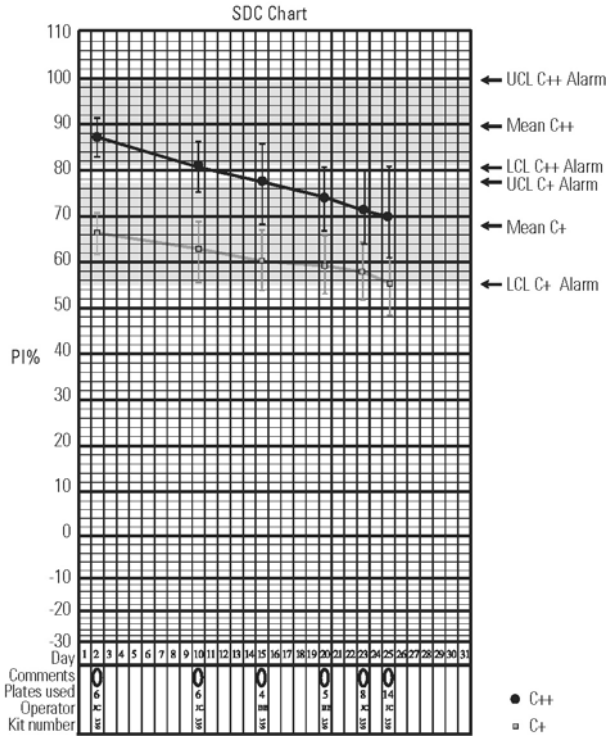
The data for each plate's controls (C++ and C+) is added (percentage inhibition). The C++ data for each plate is very similar for all tests and falls within the allowable limits of the test (grey areas). However, the C+ data shows that there is a gradual increase in values (from around 60% to 74%) which alerts users to a problem either with that control, or with the relationship of C++ to C+ in terms of optical density (OD) (which can be examined by consulting DDD charts of OD data)

Overview of producer/user/national/regional and international responsibilities

The responsibilities of the various stakeholders in relation to the sustainable supply and effective use of kits are listed in Table IV. Although the players have been separated, the areas of responsibility are inter-related and all factors have to be in place to optimise any use of kits to provide data to help in disease control.

Training

The role of training cannot be over-emphasised. Kit use without training is bound to fail. Training should concentrate on providing a fundamental understanding of the principles of disease and of the relevance of the data produced using kits for monitoring the specific analytes produced by the disease process (e.g. antibodies or antigens). Training in epidemiology is important to allow planning and assessment of the relevance of data, particularly where surveillance is involved. This is often poorly understood and the links between the epidemiologist and laboratory personnel (where different) should be strengthened to allow better planning and data management.



UCL: upper control limit
LCL: lower control limit (limits of variation given by the producer)

Fig. 10
Summary data chart summarises all plate data for two controls over time for a competitive enzyme-linked immunosorbent assay
The results show percentage inhibition (PI) values for strong positive (C++) and weaker positive (C+) control sera. The days on which the tests were performed are indicated along the bottom of the chart. The points represent the mean of the control values of the total plates run on each test (overall test variation). Each point refers to a test comprising one or more plates. The data from these individual plots can be looked at by consulting the relevant daily detailed data charts to examine variation (length of error bars). In this case both controls showed a gradual decrease over time
The LCL of the C++ is reached on the third set of tests, whereas the C+ reaches the acceptable LCL on the sixth test. The data indicate that there is loss of activity (competition) with both controls and that actions to remedy this are needed

Reference standards

Internationally accepted standards prove to be a difficult area for most diseases. A new assay may be measuring something for the first time so there is no reference preparation. Standards can be developed for use at various levels and an overview is shown in Table V. One key problem in validating tests for diagnosis and surveillance is that most categories of what could be termed international (definitive) standards are not available or are not yet agreed upon. This is partly because it is extremely difficult to produce standards that can be adapted to reflect the effects of subtle qualitative differences in all biological situations and because the maintenance and distribution of such standards require great expertise and are very expensive. In the main, 'in-house' and 'working standards' are used in biological testing. This leads to difficulties in assessing the relevance of data from a variety of test formats and complicates issues such as assessing DSn.

International organisations

Some observations can be made from an international perspective to illustrate problems with the use of kits in the context of the OIE guidelines. Such examples illustrate factors influencing the use of kits, some of which are politically motivated and illustrate bad practice. The main criticism to date has been the lack of planning for fitness for purpose and validation. This is the main reason that the OIE guidelines were formulated. The transition from research-based reagents produced in institutions where facilities are good, to kits, has always been difficult and rather ad hoc. Validation has been mainly through trial and error without adequately planned cooperative efforts. Development of the sustainable supply of kits is difficult since there is a requirement for significant resources and the veterinary market is highly fragmented. Generally, the level of quality management in diagnosis is poor in developed countries and even more so in developing countries that do not support the veterinary field. There is a danger that sampling and testing activities that yield high amounts of data are automatically seen as productive and useful, but in reality, in most control programmes much of the data produced is statistically non-viable and a waste of money. This is hard hitting, but close inspection of most data that is based on non-valid criteria would support this stance. The OIE guidelines attempt to address this situation. The success of the implementation of the guidelines will be affected by resources. Where there have been large inputs into campaigns, there is a chance that properly validated kits will be available. However, large inputs are not common, nor do planners, even those with numerous resources at their disposal, consider diagnostic and surveillance factors seriously. Even the large-scale rinderpest campaign ignored these factors and kits for measuring antibodies for diagnosis or differential diagnosis and for monitoring the efficiency of vaccination were not planned, and measures originating from a laboratory outside the planned project had to rescue this situation. The kits that they introduced have been largely instrumental in confirming the eradication of rinderpest. Foot and mouth disease (FMD) is a good example of the lack of direction in development. The use of non-structural proteins of FMD virus in the measurement of antibodies to differentiate vaccinated from infected animals has gone on for about 12 years. There are still no agreed definitive test(s) and no reference standards, there is large-scale

Table IV
Overview of responsibilities in relation to the successful use (and sustainable supply) of test kits

Participant	Responsibility	Implications
Producers	Supply robust, rugged, and complete tests in kit form Supply kits that are fit for use on leaving supplier Define fitness for purpose Produce good protocols Control standards Gather and analyse data (EQA) Respond to problems (trouble-shooting) Provide help desk services	Validation pathway criteria followed
Users	Collect and check kit Use kit/data process/report Internal quality control (continuous and transparent) Trouble-shoot/adjust/use Trouble-shoot/report failure (links with producer maintained) Train and monitor staff	Effective use of kits in control programmes. Effective staff who understand principles and practice of tests
National organisations	Train staff Monitor laboratory (EQA) Adopt standards Plan with knowledge of tests (surveys) Accreditation pathway	Tests are used efficiently in well-planned programmes
Regional organisations	Train the trainers Standardise (regional standards set and adopted) Harmonisation exercises (proficiency testing, ring-tests, etc.) Collate and report results (epidemiology) EQA	Cooperation and coordination of efforts and maintenance of standards. Generation of understanding and transparency
International organisations	Harmonise tests Set standards (e.g. World Organisation for Animal Health) Monitor test developments Stimulate test development Fund test development Fund and perform training Provide management advice	Stimulation of good practice and effective testing. Consideration of needs of all countries in managing diagnosis and surveillance leading to better control and eradication of diseases

EQA: external quality assurance

argument between different developers about the relative DSn and DSp of various kits, and there are commercial operations selling kits with ill-defined or no 'fitness for purpose'. This situation is gradually being controlled through reference to the principles of the OIE guidelines. It is hoped that suppliers will submit validation data to the OIE so that there is a transparent dossier that can be consulted by users who want to obtain the most appropriate kit for their purposes.

Cost will be a major factor in any kit supply. Validation is expensive and accelerated approaches through large-scale cooperation have to be paid for, so the final cost of the kits will be high. This has the effect of turning countries towards cheaper kits with less of a validation profile, which could be damaging to national campaigns. The expenses of validated testing have to be taken into account when drawing up plans. Ultimately, the economic advantage gained by international recognition of the absence of disease will determine whether this support is available.

It is important that there be a standardised international validation procedure to follow. The OIE guidelines are designed to establish international standards and within these, it should be possible to judge the appropriateness of tests and make recommendations as to how they can be used and further developed to meet the required standards and attain official recognition. Previously, the unleashing of tests that were not validated sufficiently, and did not meet validation criteria, has produced both apparent and more

Table V
Overview of standards for biological testing

Standards	Information
International Standard (IS) Must be used to calibrate a new method for biological analyte Such standards are of limited quantity and usually associated with calibration of national or laboratory standards or reference materials An International Unit for activity is then assigned after extensive collaboration between several different laboratories. Such standards are usually regarded as the most reliable standards	Collected-tested-aliquoted under the responsibility of the accepted world body, e.g. World Health Organization International Laboratory for Biological Standards. These are then extensively tested for potency and stability Extremely rigorous standards are employed in preparation and storage of such preparations including: – avoidance of contaminants with enzymes such as peptidases from source material – prevention of adsorption by addition of carrier substances – avoidance of oxidation by containing the samples in an atmosphere of inert gas – limitation of moisture by desiccation, storage in the dark at – 20°C
International Reference Preparation (IRP) Produced from IS	The IRP is regarded as a preparation which does not meet the demanding criteria of an IS but nonetheless is useful for method standardisation
Reference materials Not as extensively tested as IS Potency and purity data are provided by producer Valuable tools supplied by high level institutions involved in various disciplines	Such standards are very useful for substances that are: – unable to be characterised by chemical and physical means – heterogeneous materials – difficult to isolate samples – scarce or expensive samples – unstable or easily altered samples – expensive or difficult to prepare samples
In-house working standards These are preparations produced by a laboratory performing assays or acquired without any reliable potency estimates	Frequently they are calibrated against an IS
Working standards Most important standard. Constitutes basis for accuracy of a routine assay (e.g. in internal quality control)	Extensive validation and testing needed for introduction of a reference not normally necessary in the preparation of a working standard Laboratory must assume responsibility of maintaining standard's quality Larger volumes needed than reference standards

worryingly, inapparent, bad data. It is expected that the OIE guidelines will increase the quality of testing and encourage users to expect more defined kits. The responsibility of the end-user is to achieve higher levels of competence through training and to gain experience in performing tests and understanding the results in the context of different control programmes.

Testing involving polymerase chain reaction

So far this paper has concentrated on serological approaches to disease diagnosis and surveillance (the use of ELISAs). The principles of validating tests involving PCR (Fig. 11) as to fitness for purpose are the same as for serologically based tests; however, the components of the PCR require that a different emphasis be put on the various

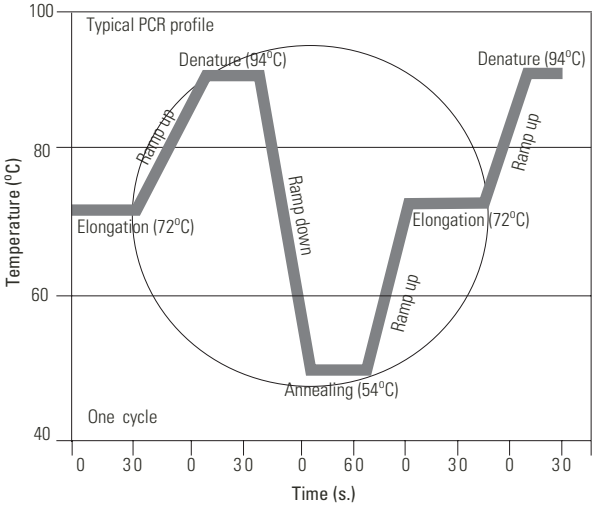


Fig. 11
Schematic illustration of a typical polymerase chain reaction temperature profile

stages of testing, in particular when considering the role of kits. The DSn is far more affected by sampling regimes and methods of nucleic extraction than serologically based tests. The PCR also suffers from technological hype where the undoubted exquisite ASn is more often than not extrapolated into an ultimate DSn without the necessary validation exercises to justify this. These two points indicate where the emphasis has to be placed in the validation pathway for PCR.

Interpretation of polymerase chain reaction results

All PCR tests are performed together with suitable controls that assist in determining the reliability of the test results obtained. A positive PCR test result indicates that the nucleic acid genome of the pathogen was present in the sample analysed. However, a positive PCR result does not necessarily indicate that the agent was present in a viable, proliferating, structurally whole or active form at the time of sample collection. A negative PCR test result indicates that the nucleic acid genome of the pathogen could not be detected under the test conditions used and suggests that the pathogen was either absent at the time of sample collection, or absent in the test material submitted, or present in sub-detectable quantities only. In a diagnostic application a negative test result should therefore not be taken as a guarantee of the absence of the pathogen within the animal (or patient) sampled. A repeat submission of appropriate sample material may, therefore, be advisable.

Note that when the recommended conditions of sample storage, submission, and preparation are not complied with, this can have a negative effect (reduced sensitivity) on the outcome of the test.

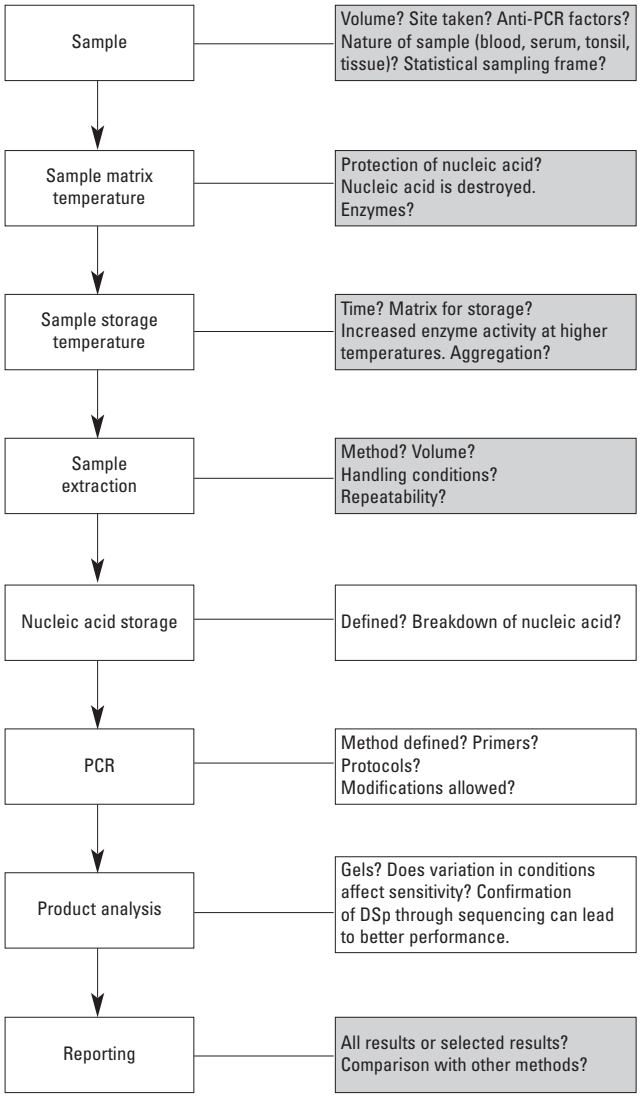


Fig. 12
Process of using polymerase chain reaction (PCR) in diagnosis
The processes of sampling and treatment/extraction are highlighted because they are of major importance in assessing the true diagnostic sensitivity of a PCR test

Validation of polymerase chain reaction

Polymerase chain reaction measures the presence or absence of a disease agent through identification and quantification of specific nucleic acid. The PCR is probably so accurate and precise because it is usually carried out using standardised precise instrumentation and because highly defined primers are widely available. Often PCR technology is examined and validated using idealised samples generated in the laboratory under experimental conditions. However, care is needed to identify a test's purpose, because more often than not it will be used for diagnosis using samples taken under field conditions. The routine use of PCR to help diagnose diseases has to be considered as a package and the DSn and DSp have to be examined and challenged using field conditions. The DSn of the PCR is far more likely than other types of assays to be affected by the sample volume, the area of the animal from which the sample was taken, and matrix and extraction methods. Generally, the equipment, diagnostic primers and protocols used in PCR are less variable than those used in serologically based assays. The use of PCR for diagnosis is illustrated in Figure 12. The factors in grey boxes must be considered more carefully when validating PCR methods. These factors are relevant to kits in the sense that specific PCR kits using defined protocols, materials and equipment could be assembled, e.g. a specific chemical to protect nucleic acid could be provided as well as nucleic acid extraction kits and control nucleic acid.

In coordinated research programmes of the Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture involving the use of PCR for diagnosing

trypanosomosis, major factors influencing DSn proved to be the use of the appropriate sample collection and extraction methods and nucleic acid storage, i.e. the way in which the samples were handled before the PCR testing. The use of primers and primer sets was less of a variable since they can be defined exactly. However, DSn is affected when protocols are tested in a wider range of countries due to slight variations in sequence of field strains, e.g. when a test developed in Europe is used in Africa. The quality of amplicons (PCR products) can be assessed through sequencing and then primers that are a better fit can be made to increase the DSn of some PCR protocols. Validation of primer sets under a variety of conditions is necessary so that DSn might be improved with reference to the poor and non-specific products which result from using inexact primers (primers with mismatches). In addition, sequencing during development can show the need to alter primers or conditions to improve the tests.

A suggested validation process has been included below, and this might be incorporated into modifications of the OIE guidelines.

Stage 1 validation

The first stage of any validation procedure involves a feasibility study to determine whether the assay can detect a range of agents (e.g. virus concentrations, virus genotypes) without background activity. This stage includes all aspects of developing and optimising the test, including the identification of the test, the determination of the target template(s), the sequence determination of the primers and the test conditions and criteria.

The first stage also involves the evaluation of the test against a panel of known positive and negative template samples to determine the ASn (which determines the smallest amount of analyte that can be detected using end-point dilution) and ASp (which determines whether there is cross-reactivity with heterologous analytes not targeted for detection). It further involves the development and standardisation of the assay in order to optimise all its components. The accuracy of the test and the repeatability of results should then be established using IQC before continuing with the validation process (Stage 2).

The following stages (2, 3 and 4) are performed less frequently or less completely because PCR methods are rarely developed into kits, although Stage 2 is the most important component of the validation process.

Stage 2 validation

This involves the determination of the DSn and DSp levels which will be used when diagnosing infection status using a relevant gold standard assay. In the case of serological assays, the OIE recommends that 300 reference samples from known infected animals and not less than 1,000 reference samples from known uninfected animals be included. It is even recommended that these values be increased to 1,000 and 5,000 respectively, to increase accuracy. In the case of PCR, these values have not been determined, but will most likely consist of significantly lower numbers. The authors propose that the PCR test is validated using field samples and field conditions in parallel to determine the DSn and DSp.

The number of samples will depend on the type of disease; however, the authors propose that using a minimum of 20 samples for rare pathogens and a minimum of 100 for more common pathogens would be an acceptable starting point to establish confidence in a test's diagnostic accuracy. The need to correlate the results with one or more other tests to add value to the PCR data is more pronounced at the early stages. As data increases, the cumulative effect will be to strengthen or weaken the argument that the PCR works 'better' or 'worse' than existing methods and these data will also help define the limits of the DSn and DSp. The data from PCR may require a rethinking of the epidemiology of some diseases, due in part to the demonstrably higher ASn of the PCR. Data from analysis of true field samples can be increased when the same method (standardised) is being applied in other laboratories. This is a very strong argument for encouraging early and sustained cooperation between laboratories in formally organised 'networks' where the tests and their ASn are harmonised using agreed reference standards. The problem of insufficient numbers of viable field samples to allow validation is common to all tests where disease agents (antigens or nucleic acid) are being directly detected. Such situations require more attention to be paid to the data generated by non gold standard statistical methods. These data are subject to strict quality controls and the additional information they provide will allow confidence factors to be defined.

Repeatability refers to the amount of agreement between replicates within or between runs. Reproducibility compares the same assay as performed between different laboratories. It is recommended that at least ten samples representing the full range of expected virus concentrations be tested in duplicate (known titres, low through high). The extent of agreement between a test value and the expected value for a sample of known virus concentration will reflect the accuracy of the assay.

Stage 3 validation

This entails the continued monitoring of the validity of assay performance in the field by calculating the predictive value of positive or negative results based on estimates of prevalence in a target animal population. This can only be done satisfactorily if DSn and DSp data are available.

Stage 4 validation

Stage 4 involves the maintenance of validation criteria using IQC. Frequent monitoring of repeatability and accuracy is needed. The OIE also recommends biannual ring-testing to determine reproducibility between laboratories, although annual testing is also described (see the following section on proficiency testing). A validated assay should consistently provide test results that identify animals as being positive or negative and accurately identify the infection status with a predetermined degree of statistical certainty.

Proficiency testing

Proficiency testing is the means used to determine the capability of a laboratory to perform the assay and effectively detect the agent (internal proficiency testing). Such testing will also contribute to ensuring that within or between laboratories performing routine diagnostic services, a specific assay is performed according to established international standards (external proficiency testing). Proficiency testing is also intended to achieve standardisation of the assays in question. This will ensure that test results obtained are reproducible within and between laboratories and will therefore give a measure of confidence that the results obtained using such an assay are reliable and trustworthy.

Internal proficiency testing

Internal proficiency testing is used to monitor the ability of laboratory personnel to produce repeatable and accurate results.

External proficiency or ring-testing

External proficiency and ring-testing are used for inter-laboratory comparisons and form part of Stages 3 and 4 of the validation process as formulated by the OIE. This can be performed on a round-robin (continuous) basis. Although it is a tedious and costly process, it is a necessary activity to achieve standardisation of PCR tests. The procedure will be that a reference laboratory periodically sends each participating laboratory (there are usually approximately three laboratories involved in the comparison) an external proficiency test, a panel of blind coded samples representing the full range of expected concentrations of the pathogen, as well as material derived from an uninfected source. Each participant then processes and tests the samples according to a particular assay method used in-house using an agreed protocol. Statistical comparisons are then made among the laboratories.

The full range of relevant pathogens that might be encountered in a clinical specimen should be tested. Samples from uninfected sources which test positive in two or three of the three laboratories can be discarded and may suggest mislabelled or contaminated samples. The goals of any clinical programme will also influence the criteria for proficiency.

Diagnosis: the prevalence paradox

Diagnosis and surveillance are aided by tests to measure the presence of a disease agent in a population by molecular methods or by the use of serological tests to detect antibodies. The OIE guidelines state that all test systems should be declared fit for purpose and that this fitness for purpose must be proved by obtaining data from studies of laboratory and field samples. Inherent in any validation study are the errors due to variations in the biological, physical and human elements. Any kit is subject to these errors and some of the effects of the errors have been discussed.

It is useful to put the errors in context of epidemiological factors since ultimately this leads to an understanding of disease in populations. This has relevance to validation since samples have to be examined from populations which are often not easy to define and this produces degrees of uncertainty in data which are not easy to quantify with the required statistical confidence limits. Most difficulties in diagnosis arise when trying to obtain data on an individual animal, i.e. to identify whether or not it has been exposed to an infectious agent.

A major problem for all diagnosticians and people developing tests is the 'prevalence paradox', where the performance of a kit is influenced by the number of animals at any time point that really have the disease, or, in more epidemiological terms, the prevalence of a disease.

The prevalence paradox

Validation studies require that the estimated DSn and DSp of a test be determined by studying populations with a known prevalence of disease; the problem is that it is not possible to know the true prevalence of disease without already having established the DSn and DSp of the test being used. Therein lies the paradox which poses major problems in the development and validation of tests.

Extremes of prevalence

Two extremes can be used to illustrate the problem as shown in Figure 13. This is an idealised situation which

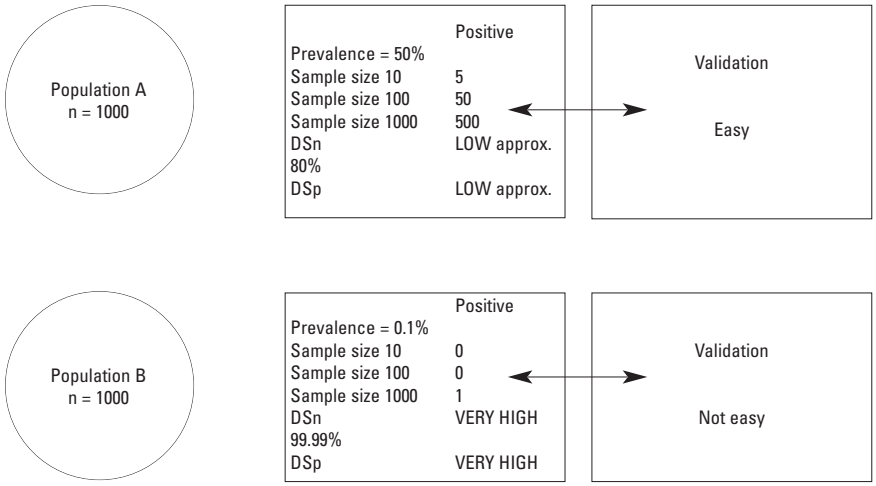


Fig. 13
Populations with very high and very low prevalences
Extremes of disease prevalence show that the needs of tests, in terms of diagnostic sensitivity (DSn), are very different. Validation in situations where there is a very low prevalence is badly affected by tests with a low DSn and populations are extremely difficult to find to act as pools for samples in validation exercises

probably never exists. Population A has a high prevalence and therefore very few samples are needed to determine that the population is infected. Population B has an extremely low prevalence, so a very high number of samples are needed to establish any disease.

The risk of misdiagnosis (false positive [FP] or false negative [FN]) is extremely high in B and therefore only tests with extreme specificity and sensitivity are acceptable. In A, the risk of a test missing a positive is very low and since most are positive, a few misdiagnosed animals will not affect the population statistic much. In terms of individuals, in B, the risk of FPs is high unless the test is almost absolute in its specificity. In A, since half of the individuals are in fact positive, the analysis that an animal is positive is correct in five out of ten cases, which means that at 80% DSp an error rate for an individual is low.

In practice, defining the disease status of a population in the field is very difficult. Various signs of infection are useful in defining an animal's status, but in a population, in time, the characteristic being measured by a test varies and infections are rarely synchronised, which means that measurements of analytes are affected quantitatively and qualitatively due to specificity factors in the test. Add to this the fact that populations are mixing, that samples are mislabelled, that people cheat and that other interventions affecting tests are not recorded, then we can see that trying to define the disease status of a population is very problematic. Without a clear understanding of the status of the herd being tested it is very difficult to statistically determine that a diagnostic test is fit for its designated purpose (particularly where there is a low prevalence of disease).

An attempt to solve the paradox is often made through laboratory experiments where animals are infected under control conditions and test formats examined with well-defined samples. The ASn and ASp of tests can be established and samples can be used to compare the results of the developing test with those of established tests and other control standards.

The DSn of a test is often determined by analysing populations in which a particular disease has never been recorded. This offers an alternative to estimating DSp but there is still the disadvantage that the populations do not always reflect the particular geographical area or breed or other field factor associated with disease. In most cases the epidemiological situation for a disease is highly variable. This depends on many factors such as disease transmissibility, geographical location, animal husbandry, population density and animal movement control. This variability can be examined and statistical confidence in data increased by repeat testing and more studies on populations, but this is expensive and needs a great deal of organisation. Figure 14 shows the relationship between DSn, DSp and prevalence. This is a rather simplified illustration of their relevance to test validation criteria. The variability of a population is shown by the width of the grey bars which relate prevalence to DSn and DSp requirements. Since there is variation, the necessary values of DSn and DSp are also variable. Most situations with regard to disease prevalence fall into the areas defined in 1 and 2, particularly the latter. In these situations a relatively wide range of DSn and DSp does not impact on the accuracy of the test. As the prevalence reduces, the DSn and DSp become far more critical for 'true' diagnosis. Such reductions in prevalence are obviously encouraging, because they suggest that progress is being made towards

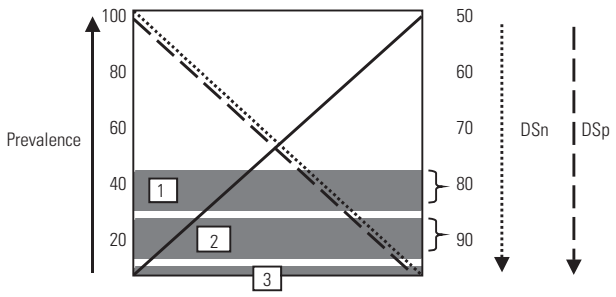


Fig. 14
Relationship of prevalence to test performance (diagnostic sensitivity [DSn] and diagnostic specificity [DSp])
The graph shows that the levels of DSn and DSp required to accurately identify the disease status of a population will vary according to the prevalence of disease. For example, in a population of 100% prevalence there is a need for only minimal (on the scale provided) Dsn, but a 100% Dsp would be needed to rule out false positive reactions. The range of 0% to 50% for DSn is more arbitrary and meant to illustrate the principle involved since any test should be better than a toss of a coin (50% correct). The zones of uncertainty for prevalence (unknown prevalence) affecting measurement of DSn are shaded in grey. The range of uncertainty for each zone is reflected by the width of the grey bands. This is the area of paradox where one determination (i.e. prevalence) relies on the other (i.e. DSn) and neither can be measured to a defined confidence. As the prevalence is reduced (real or apparent) then the need for a high DSn is more and more marked. There are few problems where the prevalence is higher than 10% (areas 1 and 2) and problems increase as the prevalence approaches and falls below 1%. The lower the prevalence the greater the DSp affects determinations of DSn

eliminating the disease. Classically, in the elimination of brucellosis it is relatively easy to achieve prevalence rates of between 5% and 0.5%. The real difficulties arise when trying to reduce the prevalence still further from 0.5% to zero, and proving it with a test(s), without undue intervention and the culling of many FP animals.

The current validation status of available kits

The validation status of most of the kits that are currently being used is strongly in question. It is probably true to say that most of them report DSn ranges of 90% to 99% and DSp ranges of 95% to 99%. This means that for most herd diagnostic purposes the majority of kits could get away with declaring themselves fit for purpose, although the criteria being used are seldom defined. However, kits are often used for studying populations where prevalences are unknown or a great deal lower than expected. In these cases, the kits are not valid and results should be viewed with suspicion. Such suspicion starts with suppliers' figures for DSn and DSp. Care must be taken to examine data to assess whether the figures tally with the studies. This is the overriding aim of the OIE guidelines. Using DSn and DSp levels that are not appropriate for the population or individual being examined is a common failing of diagnosticians.

Accumulation of data

The true status of the DSn and DSp of any test often relies on a more cumulative statistical approach where a test is used to analyse populations using a variety of sampling regimes. This in turn leads us to point out the importance of efficient data management, collection and analysis, where kits are used in a wider context to increase the amount of validation data. Data can then be from rather poorly planned sources. On analysis of data, tests can be modified to meet requirements for increased or decreased DSn and DSp. Validation is a continuous process involving data capture, analysis, modification of conditions and re-assessment of fitness for purpose. The problem with this is that tests have to be released at some point, so understanding a test's performance in all situations is impossible. The responsibility becomes the user's at this point, since he/she has the necessary reagents to evaluate the test's performance when using local samples. The laboratory can alter test values (e.g. cut-off estimates) and revise DSn and DSp characteristics within the context of planned sampling frames for a particular disease. The validation process should be strengthened by users reporting such studies to international organisations as well as to the producer, so that other users can benefit and so that producers can respond by developing new systems which meet the needs of laboratories. There is no formal arrangement for this type of development. Some companies do encourage reporting and will act on data, others do not. The level of understanding of tests and their use and flexibility are also very poor and without this there is little chance that tests can be maximised and it will be impossible to control the flow of validation data. The development of a web-based portal to retrieve data from laboratories using kits would be beneficial. This would act as a source of data to allow increased validation and would encourage laboratory personnel to participate. Problems could be solved on-line in a more public forum and users could suggest developments and agree upon protocols which would result in statistically more accurate data whose use could be extended. The FAO/IAEA Joint Division of Nuclear Techniques in Food and Agriculture is considering developing such a portal to link producers and users, and it will maintain a link with the OIE, for whom such a tool could be extremely useful in supporting their work in this area. A major feature of this portal would be to provide a focus for data storage, access to independent analysis and help, and increased data for tests to provide

higher statistical confidence. The portal would also promote good practice and cooperation between users and producers.

Conclusions

The terms 'ruggedness' and 'robustness' have been defined and their importance in the validation of tests emphasised. Robustness was defined as the resistance of a kit, or set of defined reagents, to reduced performance due to any physical factors during transport or in the laboratory. Ruggedness involved resistance to variations in laboratory technique and the effects of sampling. The responsibilities of the producer and end-user in the supply and use of kits were examined and helpful advice provided to improve methods of testing and continuously monitoring kits, such as charting methods to record IQC data, trouble-shooting problems with kits and reporting problems. The end-user should be trained enough to be able to determine which parts of a kit are not functioning and take steps to eliminate the problem or at least report findings for action by the supplier. End-users should also have the skills to adapt tests to local conditions and produce quality-controlled data to justify findings. In turn, the supplier should take steps to gather data continuously and act on any problems. This is seen as part of validation. The authors examined the problems posed by the prevalence paradox, whereby the prevalence among populations used for validation is variable and so 'true' DSn is not easy to measure with any confidence. The importance of obtaining and collating data from well-run tests to increase the confidence in validation was stressed. A key factor in increasing test quality is agreement on the production and characterisation of reference samples. This is not an easy task due to the particular problems of variation seen in biological systems; the heterogeneous nature of analytes; the variation in needs for hard-to-define populations and the cost of maintaining and distributing materials. International organisations have, so far, played a major role in facilitating the development of kits to be used in disease control by supplying training to allow the materials to be used efficiently and emphasising the importance of quality control when obtaining data. This process suffers from a discontinuity of funding and poor coordination of planning in the medium and long term. Diagnostic and surveillance activities are often forgotten in the planning process at national and certainly regional levels. If followed, the OIE guidelines, possibly with modifications concerning PCR testing, would solve many of the problems associated with the supply of kits used in diagnosis and surveillance. An overview of issues surrounding validation is shown in the literature review (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16).

Validation des kits de diagnostic et de surveillance des maladies du bétail : responsabilités des producteurs et des utilisateurs finaux

J.R. Crowther, H. Unger & G.J. Viljoen

Résumé

La Division mixte de l'Organisation pour l'alimentation et l'agriculture/Agence internationale de l'énergie atomique (FAO/AIEA) des techniques nucléaires dans l'alimentation et l'agriculture basée à l'AIEA à Vienne a une vaste expérience en matière d'aide à l'élaboration et à la validation des tests de diagnostic et a apporté un appui précieux à l'Organisation mondiale de la santé animale (OIE) pour l'élaboration des normes. Le présent article est axé sur la méthode de dosage immuno-enzymatique et l'amplification en chaîne par la polymérase qui sont les principales techniques appliquées pour le diagnostic et la surveillance des maladies. Les problèmes posés par la terminologie et les facteurs intervenant dans la production, l'approvisionnement et la validation des kits sont examinés, en mettant tout particulièrement l'accent sur la robustesse et la stabilité des tests. Les auteurs passent en revue les responsabilités des

différentes parties prenantes (producteurs, distributeurs, utilisateurs et organisations nationales/internationales) en matière d'obtention de données ayant fait l'objet d'un contrôle de qualité pour résoudre les problèmes de diagnostic et de surveillance. Le rôle du contrôle interne de qualité (contrôles internes de compétences) et de l'assurance externe de qualité (contrôles externes de compétences) ainsi que les outils permettant de trouver des solutions aux problèmes posés par les kits sont examinés.

Mots-clés
Amplification en chaîne par polymérase – Diagnostic – Kit – Ligne directrice – Méthode de dosage immuno-enzymatique – Organisation mondiale de la santé animale (OIE) – Robustesse – Sensibilité diagnostique – Spécificité diagnostique – Stabilité – Surveillance – Validation.

Responsabilidades de fabricantes y usuarios finales en la validación de *kits* de diagnóstico y vigilancia de las enfermedades del ganado

J.R. Crowther, H. Unger & G.J. Viljoen

Resumen
La División de Técnicas Nucleares en la Alimentación y la Agricultura es una unidad mixta de la FAO (Organización de las Naciones Unidas para la Agricultura y la Alimentación) y el OIEA (Organismo Internacional de Energía Atómica) radicada en la sede del OIEA, Viena (Austria), que goza de amplia experiencia en actividades de ayuda a la elaboración y validación de ensayos y ha prestado un gran apoyo a la formulación de las normas de la Organización Mundial de Sanidad Animal (OIE). Los autores se centran en el ensayo inmunoenzimático (ELISA) y la reacción en cadena de la polimerasa (PCR), que son las dos principales técnicas utilizadas en labores de diagnóstico y vigilancia. Ante todo exponen los problemas ligados a la terminología y a los factores que intervienen en la fabricación, el suministro y la validación de los *kits*, haciendo especial hincapié en la importancia de la fiabilidad (*robustness*) y la resistencia a la variación de factores externos (*ruggedness*) de las pruebas. Después analizan las responsabilidades de las distintas partes (fabricantes, distribuidores, usuarios y organizaciones nacionales o internacionales) a la hora de generar datos sometidos a un control de calidad para resolver problemas de diagnóstico y vigilancia. Por último, examinan las funciones del control de calidad (pruebas de rendimiento internas) y la garantía de calidad (pruebas de rendimiento externas), así como las posibles ayudas para resolver problemas relacionados con los *kits*.

Palabras clave
Diagnóstico – Directriz – Ensayo inmunoenzimático – Especificidad de diagnóstico – Fiabilidad – Kit – Organización Mundial de Sanidad Animal (OIE) – Reacción en cadena de la polimerasa – Resistencia a la variación de factores externos – Sensibilidad de diagnóstico – Validación – Vigilancia.

References

1. Crowther J.R. (2000). – Charting methods for internal quality control. *In* The ELISA guidebook: methods in molecular biology, Chapter 9. Humana Press, Totowa, New Jersey, 347-394.

2. Crowther J.R. (2000). – Validation of diagnostic tests for infectious diseases. *In* The ELISA guidebook: methods in molecular biology, Chapter 8. Humana Press, Totowa, New Jersey, 301-345.

3. Enoe C., Georgiadis M.P. & Johnson W.O. (2000). – Estimating the sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. vet. Med.*, **45** (1-2), 61-81.

4. Greiner M. & Gardner I.A. (2000). – Application of diagnostic tests in veterinary epidemiologic studies. *Prev. vet. Med.*, **45** (1-2), 43-59.

5. Greiner M. & Gardner I.A. (2000). – Epidemiologic issues in the validation of veterinary diagnostic tests. *Prev. vet. Med.*, **45** (1-2), 3-22.

6. Greiner M., Pfeiffer D. & Smith R.D. (2000). – Principles and practical application of the receiver operating characteristic (ROC) analysis for diagnostic tests. *Prev. vet. Med.*, **45** (1-2), 23-41.

7. Jacobson R.H. (1998). – Validation of serological assays for diagnosis of infectious diseases. *In* Veterinary laboratories for infectious diseases. *Rev. sci. tech. Off. int. Epiz.*, **17** (2), 469-486.

8. World Organisation for Animal Health (OIE) (2002). – OIE Guide 3: laboratory proficiency testing. *In* OIE Quality standard and guidelines for veterinary laboratories: infectious diseases. OIE, Paris, 53-63.

9. World Organisation for Animal Health (OIE) (2002). – OIE Standard for management and technical requirements for laboratories conducting tests for infectious animal diseases. *In* OIE quality standard and guidelines for veterinary laboratories: infectious diseases. OIE, Paris, 1-31.

10. World Organisation for Animal Health (OIE) (2004). – Principles of validation of diagnostic assays for infectious diseases. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 5th Ed. Volume I, Part 1, Chapter I.1.3. OIE, Paris, 21-29.

11. World Organisation for Animal Health (OIE) (2004). – Quality management in veterinary testing laboratories. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 5th Ed. Volume I, Part 1, Chapter I.1.2. OIE, Paris, 14-20.

12. World Organisation for Animal Health (OIE) (2004). – Sampling methods. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 5th Ed. Volume I, Part 1, Chapter I.1.1. OIE, Paris, 3-13.

13. World Organisation for Animal Health (OIE) (2004). – Validation and quality control of polymerase chain reaction methods used for the diagnosis of infectious diseases. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 5th Ed. Volume I, Part 1, Chapter I.1.4. OIE, Paris, 30-36.

14. Wright P.F. (1998). – International standards for test methods and reference sera for diagnostic tests for antibody detection. *In* Veterinary laboratories for infectious diseases. *Rev. sci. tech. Off. int. Epiz.*, **17** (2), 527-533.

15. 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.

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