CHAPTER 2.2.8.

COMPARABILITY OF ASSAYS AFTER CHANGES IN A VALIDATED TEST METHOD

INTRODUCTION

There are many reasons why validated assays undergo changes over time. Replacement of depleted reagents is probably the most common example of a minor change to a validated assay (see Chapter 1.1.6 Validation of diagnostic assays for infectious diseases of terrestrial animals, *Figure 1, Assay development and validation pathways). Small changes may be driven by: the availability of less expensive or better reagents, e.g. nucleic acid extraction reagents for molecular tests; the need for improved standardisation, e.g. changing from operator-coated to pre-coated enzyme-linked immunosorbent assay (ELISA) plates; increased throughput requirements, e.g. manual versus robotic handling, etc. Such minor changes usually require an experimental study to assess whether the performance characteristics of the validated assay are still comparable with the new procedure (Table 1, Figure 1). There are other important variables outside of the assay that may require verification, such as the nature of the target population, the species and the specimen. For example, experienced laboratory diagnosticians would be cautious about applying a competitive brucellosis antibody ELISA to cattle in Latin America if it had been validated specifically for cattle in Canada (Gall* et al., *1998). Assays are often applied to a species other than that for which they had been originally validated, e.g. domestic chickens versus wild birds or beef versus dairy cattle. Other changes included the use of different test specimens, e.g. tracheal swabs versus cloacal swabs from birds for diagnosis of avian influenza using molecular assays. Under these circumstances, a verification study would be necessary to validate the performance characteristics of the test under the new circumstances.*

Controversy exists about what constitutes a "minor" and a "major" change for a diagnostic assay. There are some changes that are regarded as major because the biological basis of the assay is fundamentally altered, for example, evolutionary changes or mutations in the nucleic acid make-up of a pathogen will require adjustments to be primers and probes. Similarly, a change from an indirect to a competitive ELISA format using a highly specific monoclonal antibody is considered a major change that would warrant complete re-validation of the assay. Table 1 provides some examples of minor and major changes that are found frequently in the use of antibody and nucleic acid detection tests. Rigorous and well designed comparability studies provide an objective assessment of whether the assay, when used with a minor change, is as comparable to, and fit for the intended use, as the validated assay. The outcome of the experimental study will determine whether or not the candidate assay requires full re-validation and consequently whether or not it can be used with confidence.

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

Verification represents evidence that the performance characteristics, e.g. accuracy and precision of a validated assay, are comparable when used in another laboratory.

Comparability is the preferred term when performance characteristics of a new test, which has undergone a minor change, are as good as those of a validated test within statistically defined limits.

Equivalence or equivalency has historically been used in some diagnostic laboratories for comparability studies. However, the term implies a more stringent requirement than fitness for intended use and also has a specific statistical meaning. For these reasons, the term is not used in this chapter.

A. SETTING UP COMPARISON EXPERIMENTS

Table 1. Examples of change in diagnostic tests

Type of change	Changes in assay	Changes in target population or specimens
Minor	Replacement of depleted reagent, e.g. positive control sample, new batch of antigen, plates, conjugate (ELISA) Change of instrument/platform, e.g. ELISA reader, incubator/shaker, thermocycler (PCR) Change from individually coated to pre-coated ELISA plates Change from manual to robotic handling (ELISA, NAD) Change in nucleic acid extraction procedure (NAD) Use of modified primer(s) or probe (partial substitution of sequences, e.g. degeneracy) Modified PCR reaction conditions using the same primers(s) and probe(s) Addition of an extra probe within the amplified region Change in probe chemistry (NAD)	
Major	Substitution of a recombinant antigen for a cell culture-derived antigen in an ELISA Change from an indirect to a competitive ELISA using a specific monoclonal antibody Change of primers and probes for different targets in different regions of the same or different gens (NAD)	Different species, e.g. cattle versus buffalo, domestic chicken versus wild birds Different specimen types, e.g. tracheal versus cloacal swabs, blood versus semen, different tissues or organs

ELISA = enzyme-linked immunosorbent assay; PCR = polymerase chain reaction; NAD = nucleic acid detection (tests);

When setting up a comparison experiment, the procedure should be guided by the purpose of the assay (Figure 1, step 1). For example, screening assays require high diagnostic sensitivity, and it is important to compare the limit of detection. In such a case a suitable dilution range and the number of replicates of a control sample have to be determined. A sufficient quantity of well characterised control sample material needs to be produced, aliquoted and stored appropriately.

If the objective is to assess and compare repeatability it is necessary to run well characterised replicates of control samples of different analyte concentrations spanning the expected operating range of the assay, e.g. a high, medium and low analyte concentration. For practical purposes, the example given in this chapter only uses a weak-positive control sample. It is good practice to visually inspect correlation of results between both methods. For example, a scatter diagram and a histogram are easily performed and provide immediate information about the type of correlation and distribution of data (Figures 2 and 3). Basic statistics help to set upper and lower limits and evaluate results, for example for limit of detection or repeatability (Tables 2 and 3). A more sophisticated approach to comparing and analysing results from a comparison study is a Bland–Altman plot (Figure 4 and Table 4).

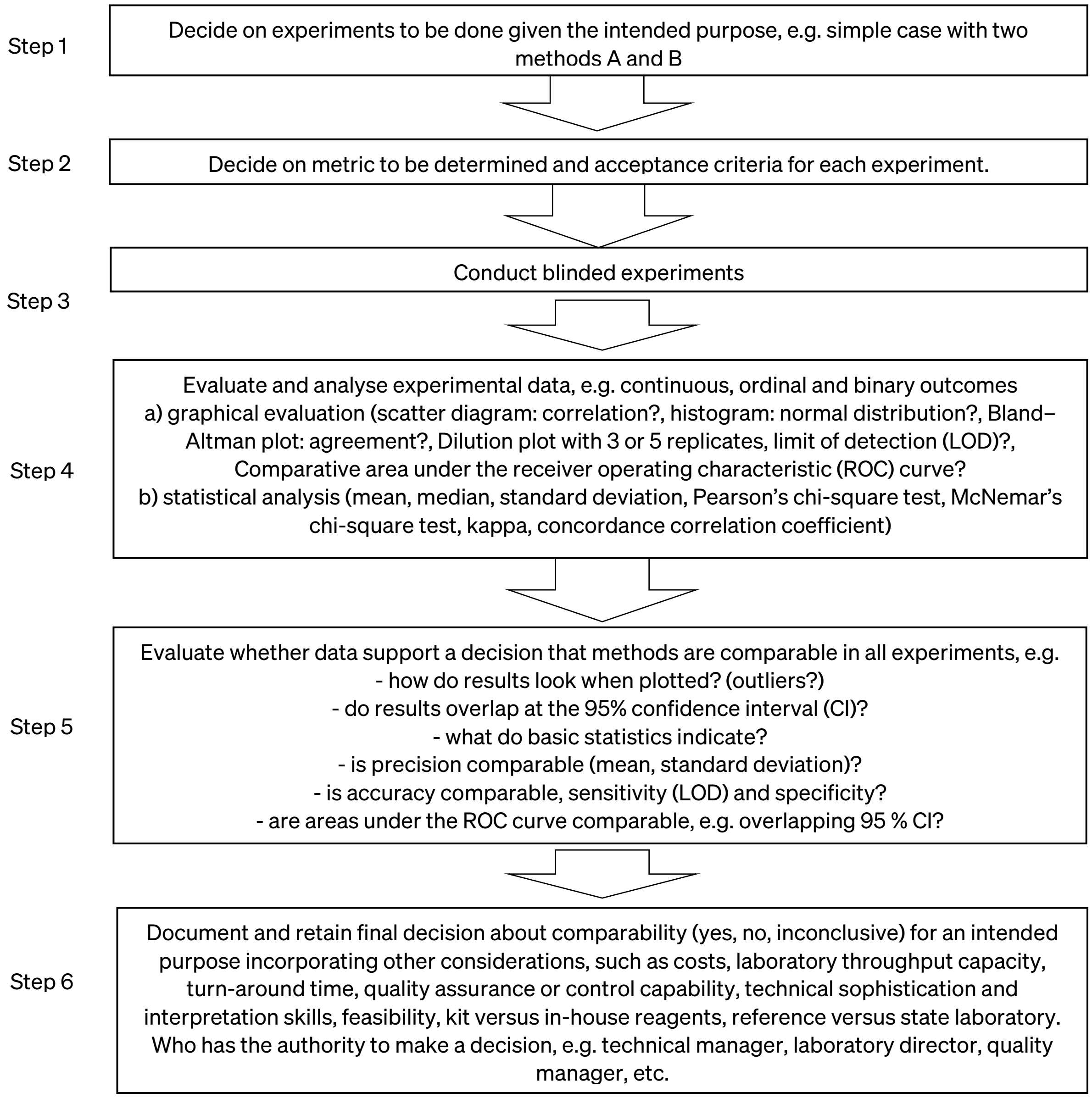
Figure 5 is an example of how to increase the efficiency of comparison experiments by assessing different parameters simultaneously on a single plate, e.g. assessing analytical sensitivity using dilution steps of a target analyte in triplicate, followed by assessing analytical specificity using a number of negative samples (which do not carry the target analyte), and finally assessing analytical sensitivity using a number of samples from infected animals with different analyte concentrations. The use of replicates in the same run and between runs allows estimation of repeatability (Figure 5 and 6).

It is important to agree on acceptance criteria to evaluate the outcome of the experiment (Figure 1, step 2), e.g. confidence intervals can be asymmetric to allow for better performance of a novel assay. In this chapter a conservative 95% confidence estimate is regarded as acceptable for a limit of detection experiment (Figures 6 and 7). For comparison of repeatability, the mean and standard deviation (SD) or direct test results plus and minus a given range can be used as acceptance criteria (Tables 3 and 4 and Figure 4). Results from a panel of infected and non-infected individuals provide information about comparative diagnostic sensitivity and specificity (Figure 8 and Tables 5 and 6).

Data used in Figures 2, 3 and 4 and Tables 2, 3 and 4 were produced using results from a repeatedly tested weak positive control sample in two TaqMan assays that target the M (M1 assay) and N (N1 assay) genes of Hendra virus.

Data for LOD experiments and plate layout in Figure 5, 6 and 7 are fictive. Data for ROC curves in Figure 8 and Tables 5 and 6 are taken from comparison experiments of different Influenza ELISAs in pigs.

Fig. 1. Factors to be considered for comparability studies of diagnostic tests.



Step 1: Decide on experiments to be done given the intended purpose, e.g. simple case with two methods A and B

Step 2: Decide on metric to be determined and acceptance criteria for each experiment.

Step 3: Conduct blinded experiments

Step 4: Evaluate and analyse experimental data, e.g. continuous, ordinal and binary outcomes
a) graphical evaluation (scatter diagram: correlation?, histogram: normal distribution?, Bland–Altman plot: agreement?, Dilution plot with 3 or 5 replicates, limit of detection (LOD)?, Comparative area under the receiver operating characteristic (ROC) curve?
b) statistical analysis (mean, median, standard deviation, Pearson's chi-square test, McNemar's chi-square test, kappa, concordance correlation coefficient)

Step 5: Evaluate whether data support a decision that methods are comparable in all experiments, e.g.
- how do results look when plotted? (outliers?)
- do results overlap at the 95% confidence interval (CI)?
- what do basic statistics indicate?
- is precision comparable (mean, standard deviation)?
- is accuracy comparable, sensitivity (LOD) and specificity?
- are areas under the ROC curve comparable, e.g. overlapping 95 % CI?

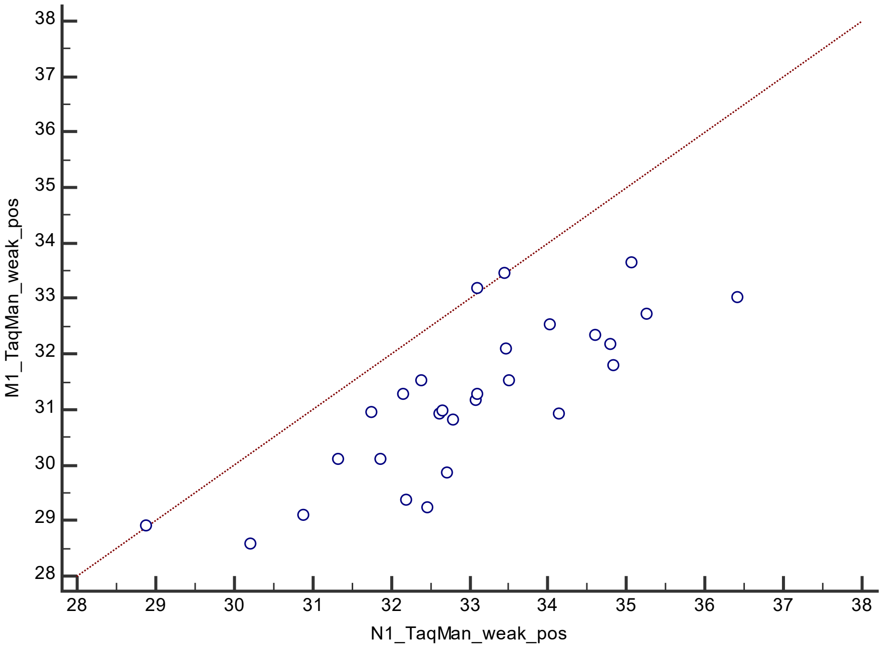
Step 6: Document and retain final decision about comparability (yes, no, inconclusive) for an intended purpose incorporating other considerations, such as costs, laboratory throughput capacity, turn-around time, quality assurance or control capability, technical sophistication and interpretation skills, feasibility, kit versus in-house reagents, reference versus state laboratory. Who has the authority to make a decision, e.g. technical manager, laboratory director, quality manager, etc.

This chapter provides an overview of different approaches to the design of experiments and interpretation of results from assay comparison studies, e.g. analytical (limit of detection) and diagnostic sensitivity, analytical and diagnostic specificity, and repeatability. There are many fundamentally different tests but the examples provided stem from or refer to experiments with nucleic acid detection (NAD) and enzyme-linked immunosorbent assays (ELISAs). It can be assumed that the principles provided in this chapter are equally applicable to other tests.

B. VISUAL INSPECTION

A **Scatter diagram** is useful for visually evaluating the correlation between both methods initially, e.g. is the relationship linear or logarithmic? Are there outliers or missing values or artefacts? The example in Figure 2 uses results from a repeatedly tested weak positive control sample from two TaqMan assays that target the M (M1 assay) and N (N1 assay) genes of Hendra virus. Results show that data for the N1 assay are shifted to the lower right compared with the M1 assay, which indicates consistently higher values for N1 than for M1 assay. However the scatter diagram does not provide information about the agreement of the two tests. There is one exception to this rule, if all results for both tests would fall along the 45% diagonal line, the agreement would be 100%. In Figure 2 only three results fall along the diagonal line. In this paper we define agreement as a set of values of a candidate test that fall within the 95% CI of the results of the established test after repeated runs of the same well characterised control sample. Correlation measures the strength of the relation between measurements from these tests and is expressed as p value.

Fig. 2. Scatter diagram of a weak positive control sample after being tested 28 times in two different Hendra TaqMan assays, M1 and N1 (results expressed as cycle threshold [Ct] values).



Further analysis of results is given in Table 2 below, where r (correlation coefficient) = 0.8 and indicates a strong, positive correlation between the two methods. *p* < 0.0001 indicates that the probability that this association is due to chance is very low and the 95%CI (confidence interval) indicates that when these methods are used on a similar subject and under similar conditions, we are 95% confident that the true unknown value of r lies between 0.61 and 0.9.

Table 2. Statistical analysis of a weak positive control sample after being tested 28 times in two Hendra TaqMan assays, M1 and N1

Variable Y	M1 TaqMan weak positive
Variable X	N1 TaqMan weak positive
Sample size	28
Correlation coefficient r	0.8015
Significance level	*p*<0.0001
95% confidence interval for r	0.6112 to 0.9042

A **histogram**, where skewing to the left or to the right or other relevant characteristics such as a bimodal distribution would be detected is shown Figure 3.

Fig. 3. Histogram of a weak positive control sample after being tested 28 times in two Hendra TaqMan assays, M1 and N1 (results expressed as cycle threshold [Ct] values).

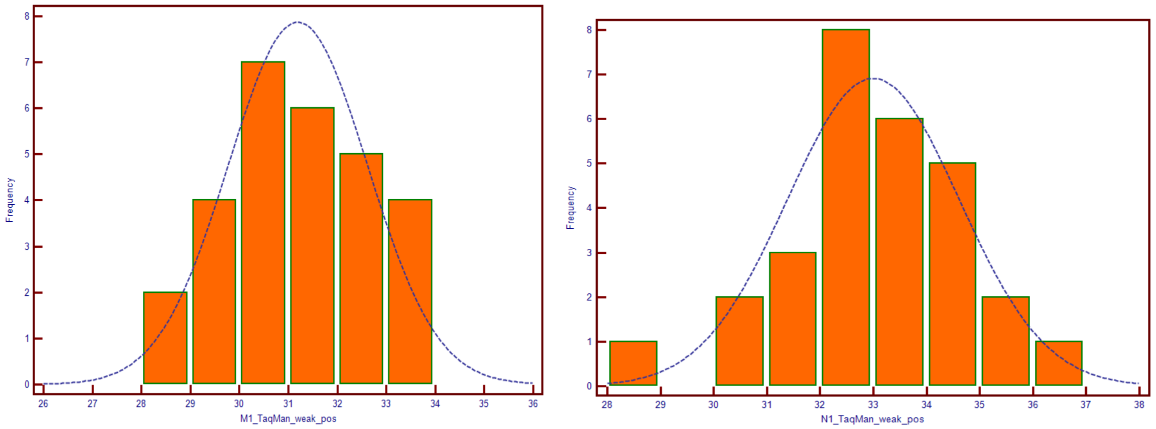


Table 3. Statistical analysis of a weak positive control sample after being tested 28 times in two Hendra TaqMan assays, M1 and N1 (results expressed as cycle threshold [Ct] values)

	M1 Taqman weak positive	**N1 Taqman weak positive**
Sample size	28	28
Lowest value	28.58	28.88
Highest value	33.63	36.42
Mean	31.19	33.00
Median	31.22	32.94
Standard deviation (SD)	1.42	1.62

C. REPEATABILITY

Assay variation can be assessed using replicates of an internal control sample when used in sequential runs over time. In this example, repeatability was compared for two different Hendra TaqMan assays targeting the N and M genes using a weak-positive internal control sample after 28 runs by the same operator on 14 days and during an 18-day period. Cycle threshold (Ct) results were summarised as the mean and standard deviation (SD) in Table 3.

Because the estimates are based on a single control sample, no formal comparison is necessary. Rather, if the predefined acceptance criteria for both assays were, e.g. 2 to 3 SDs or ± 2 to 3 Ct values then both assays would be considered comparable.

D. BLAND–ALTMAN PLOT

A very efficient way to simultaneously display and analyse results from a comparison study where 2 different measurements are made on the each sample (so called matched-pairs design) is a Tukey mean difference plot (Bland & Altman, 1999, 2007; Kozk & Wnuk, 2014). The plot is useful for revealing a relationship between the differences and the averages, evaluating any systematic biases and identifying possible outliers. Average Ct values obtained with methods A (M1 TaqMan assay) and B (N1 TaqMan assay) across a range of results are plotted along the x-axis and differences in the mean values, e.g. A minus B, are displayed on the y-axis (Figure 4, Table 4). In the example, the M1 TaqMan is compared with N1 TaqMan using results from a weak positive control sample after 28 runs. All differences are negative because the mean values for N1 TaqMan (Ct 33) are higher than M1

TaqMan (31.99) (Table 3). Subtracting N1 from M1 reveals that almost all values are negative, which indicates a systematic bias, e.g. the average difference between N1 and M1 was –1.81 (bias). Horizontal lines are drawn at the mean difference (–1.81), and at the limits of agreement, which are defined as the mean difference plus and minus 1.96 times the standard deviation of the differences, which, expressed in average Ct differences, are 0.1 to –3.71. Results in Table 4 show that the 95% CI for the mean difference (–2.18 to –1.43) exclude zero and hence, it cannot be concluded that the two methods are comparable. Adjustment of the cut-off could help to counterbalance this apparent systematic bias between both methods after corroborating these findings with larger sample numbers and over the range of expected results.

One measurement is made by each method on each sample (matched-pairs design) to compare repeatability between the two methods over the measurement range, e.g. for competitive antibody ELISAs it is known that variation increases with the decrease of the analyte concentration in the sample. A negative or weak-positive control sample may have significantly higher variation than high-positive control samples. Figure 4 shows three results that lie along the line for the 0 value, e.g. the line where results for both tests are identical because the difference is zero. These are the same results that fall on the diagonal line of equality in the scatter diagram in Figure 2 at values of 28.9, 33.10 and 33.45 for both assays.

Fig. 4. Bland–Altman plot showing differences of cycle threshold (Ct) values in two Hendra TaqMan assays, M1 and N1 for a weak-positive control sample after being tested 28 times.

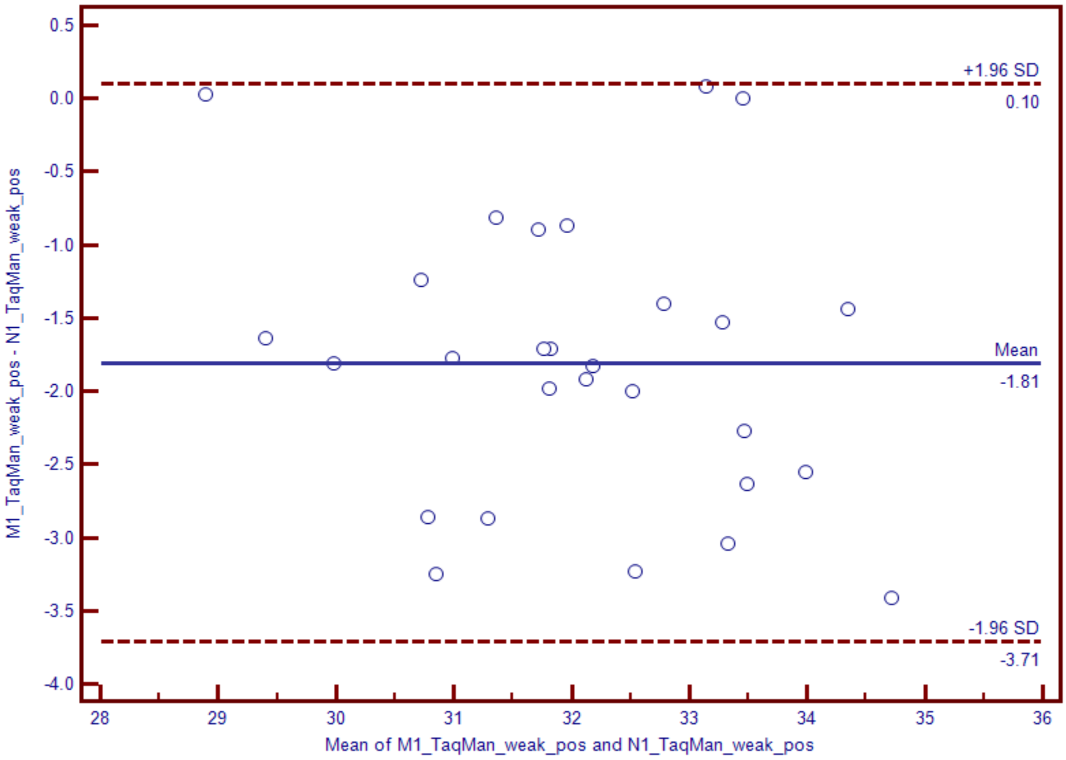


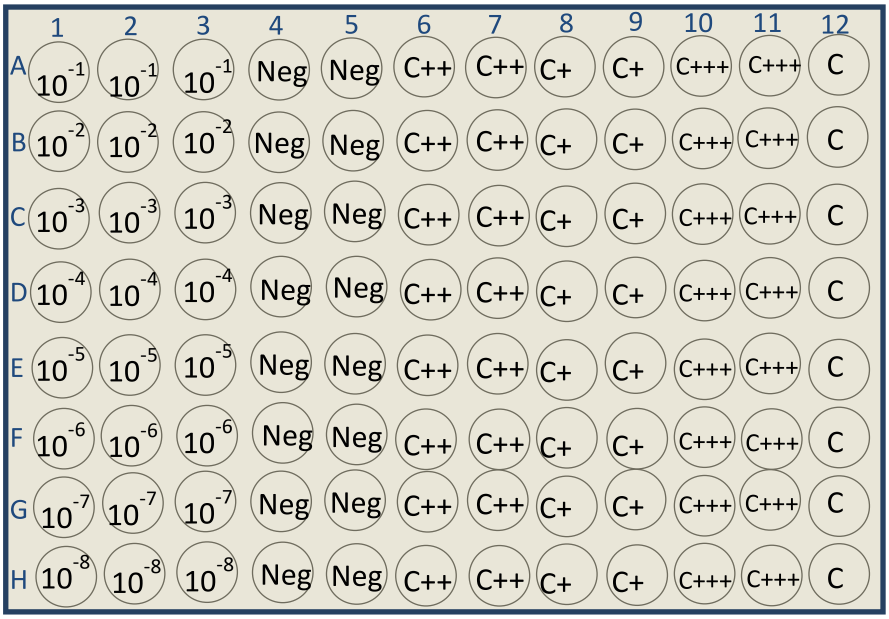
Table 4. Statistical analysis for Bland–Altman plot

Method A	M1 TaqMan weak positive
Method B	N1 TaqMan weak positive
Differences	
Sample size	28
Arithmetic mean	–1.8054
95% CI	–2.1831 to –1.4276
Standard deviation	0.9741

E. LIMIT OF DETECTION (LOD) EXPERIMENT

An example of a plate layout for a comparability study of a molecular test is given in Figure 5. Limit of detection (three replicates of positive samples in a dilution series from 10^{-1} to 10^{-8} [analytical sensitivity]), diagnostic specificity (negative diagnostic samples from non-infected animals or animals that have been infected with a non-target pathogen [Neg] tested in duplicate), diagnostic sensitivity (samples from field infected animals of different activity, e.g. extremely high positive [C+++], very high positive [C++], high positive [C+] and positive [C] tested in duplicate), and repeatability are assessed. Cross-contamination can be also evaluated as strong positive samples are placed next to negative samples.

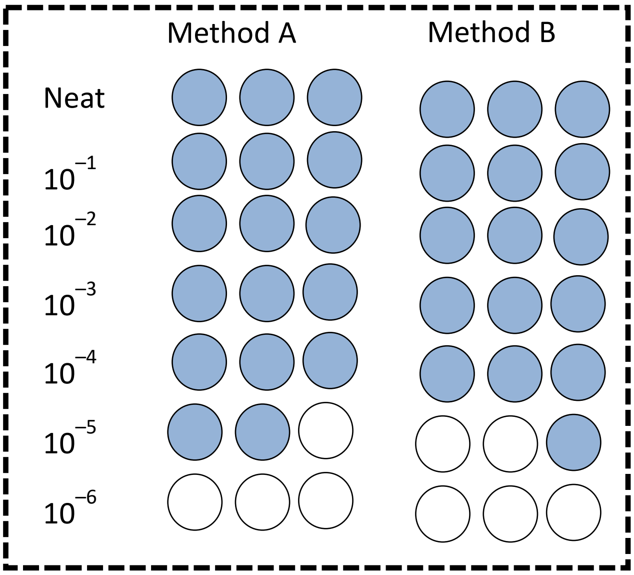
Fig. 5. Layout of 96-well plate to assess analytical Se, diagnostic Se and Sp and repeatability.



	1	2	3	4	5	6	7	8	9	10	11	12
A	10^{-1}	10^{-1}	10^{-1}	Neg	Neg	C++	C++	C+	C+	C+++	C+++	C
B	10^{-2}	10^{-2}	10^{-2}	Neg	Neg	C++	C++	C+	C+	C+++	C+++	C
C	10^{-3}	10^{-3}	10^{-3}	Neg	Neg	C++	C++	C+	C+	C+++	C+++	C
D	10^{-4}	10^{-4}	10^{-4}	Neg	Neg	C++	C++	C+	C+	C+++	C+++	C
E	10^{-5}	10^{-5}	10^{-5}	Neg	Neg	C++	C++	C+	C+	C+++	C+++	C
F	10^{-6}	10^{-6}	10^{-6}	Neg	Neg	C++	C++	C+	C+	C+++	C+++	C
G	10^{-7}	10^{-7}	10^{-7}	Neg	Neg	C++	C++	C+	C+	C+++	C+++	C
H	10^{-8}	10^{-8}	10^{-8}	Neg	Neg	C++	C++	C+	C+	C+++	C+++	C

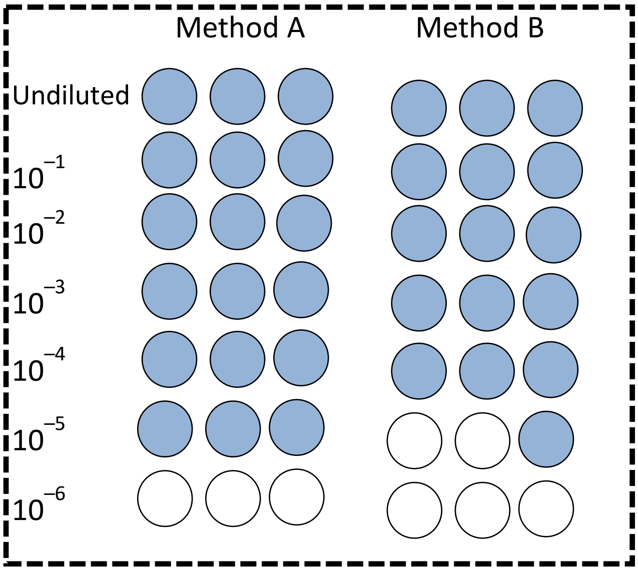
The limit of detection (LOD) is a measure of the analytical sensitivity (ASe) of an assay. The LOD is the estimated amount of analyte in a specified matrix that would produce a positive result at least a specified per cent of the time. Figures 6 and 7 represent hypothetical results of a LOD experiment. For example, in a titration using tenfold dilutions all replicates at all dilutions might show either 100% or 0% response. There are two choices at that point. The last dilution showing 100% response may be accepted as a conservative estimate of the lower limit of detection. A more accurate estimate may be obtained by a second stage experiment using narrower intervals in the dilution scheme focusing on the region between 100% and 0%. The first step is to produce, aliquot and blind a sufficient number of samples to carry out the experiment. The second step is to produce a set of analyte dilutions, preferably using sample matrix as diluent, rather than buffer, which reflect the measurement range of the method, e.g. in a tenfold dilution series, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}. Practical examples often use 3–5 replicates per dilution step. Using a conservative estimate of 100% (all 3 wells at the highest dilution dark = positive) Figures 6 and 7 show comparable and non-comparable outcomes, respectively. Method A represents the validated and method B the new method.

Fig. 6. Example of a limit of detection (LOD) experiment with an acceptable outcome.



In Figure 6 at the 10^{-4} dilution all replicates from method A and method B are positive (blue). At the 10^{-5} dilution only two out of three wells are positive for method A and one out of three is positive for method B. As the limit of detection is defined as the dilution where all wells must be positive results from the 10^{-5} dilution downwards are not considered for comparability. Applying these criteria the limit of detection in Figure 6 is the same for methods A and B and therefore the two methods can be regarded as comparable.

Fig. 7. Example of a limit of detection (LOD) experiment with a non-acceptable outcome.



In Figure 7 the highest dilution of all three replicates is at 10^{-5} for method A. In contrast the highest dilution where all three replicates are still positive for method B is at 10^{-4}. Consequently, the limits of detection of methods A and B cannot be considered comparable if a single log dilution is not acceptable. It is advisable to repeat the experiment several times before making a final decision on comparability.

F. COMPARISON OF ROC CURVES

Receiver operating characteristic (ROC) analysis is a powerful method to assess and compare the overall accuracy of a diagnostic test, e.g. diagnostic sensitivity (DSe) and diagnostic specificity (DSp) at different cut-offs of one or more different or modified tests (Greiner *et al.,* 2000). The central measurement is the area under the curve (AUC), e.g. a value of 1 indicates a test with 100% DSe and 100% DSp. In this case there is perfect separation of the values of the two groups, i.e. there is no overlapping of the distributions and the ROC curve will reach the upper left corner of the plot. In contrast a value of 0.5 indicates no discrimination between infected and non-infected individuals beyond chance; the ROC curve coincides with the diagonal indicating that the test is useless. Values between 0.5 and ≤0.7 can be considered less accurate, from 0.7 to ≤0.9 moderately accurate and from 0.9 to <1 as highly accurate (Greiner *et al.,* 2000).

Figure 8 shows results from four antibody ELISAs for influenza virus in pigs. The serum panel consisted of 168 positive and 77 negative sera using the haemaglutination inhibition as a reference test (*n*=245). When compared on these diagnostic samples the following ranking was established: test 1 (AUC = 0.988), test 4 (AUC = 0.988), test 3 (AUC = 0.929) and test 2 (AUC = 0.814) (Table 5). For this matched pair design experiment the 95% CI for the differences of the Area Under the ROC curve (AUC) can be used as an indicator for statistical significance (Table 6). In summary, the test with the best DSe and DSp and the highest area under the curve (0.988) is test 1. The result of 0.988 means that a randomly selected individual from the positive group has a test value larger than that for a randomly chosen individual from the negative group 99% of the time (Zweig & Campbell, 1993).

Another way to compare results is to evaluate AUC differences. For example for test 1 and 2 the difference between the AUC was 0.0173, for test 1 and test 3 the result was 0.058, for test 1 and 4 the result was 0.00019, for test 2 and 3 the result was 0.1150, for test 2 and 4 the result was 0.174 and for test 3 and 4 the result was 0.0586 (Table 6). The tests with the highest AUCs and smallest difference of AUCs were test 1 and test 4, e.g. both tests had an AUC of 0.988, overlapping 95% CI, the difference between the AUCs was as low as 0.00019 and a *p* value of 0.98 indicated no statistically significant difference at the 5% significance level. Lower AUC values, lack of overlapping 95%CI, significantly increased AUCs and *p* values <0.05 indicated lack of agreement of other test combinations, e.g. 1 vs 3, 1 vs 2, 2 vs 3, 2 vs 4 and 3 vs 4. Test 1 0.964 to 0.997, test 2 0.760 to 0.861, test 3 0.889 to 0.958, test 4 0.965 to 0.997.

More complex comparison studies with tests based on similar diagnostic and biological principles using frequentist and classical statistical approaches have been published (Brocchi *et al.,* 2006; Engel *et al.,* 2008). See also Chapter 2.2.5. *Statistical approaches to validation*.

Fig. 8. Comparison of receiver operating characteristic (ROC) area under the curve (AUC) for four different ELISAs to detect antibodies against influenza virus in pigs.

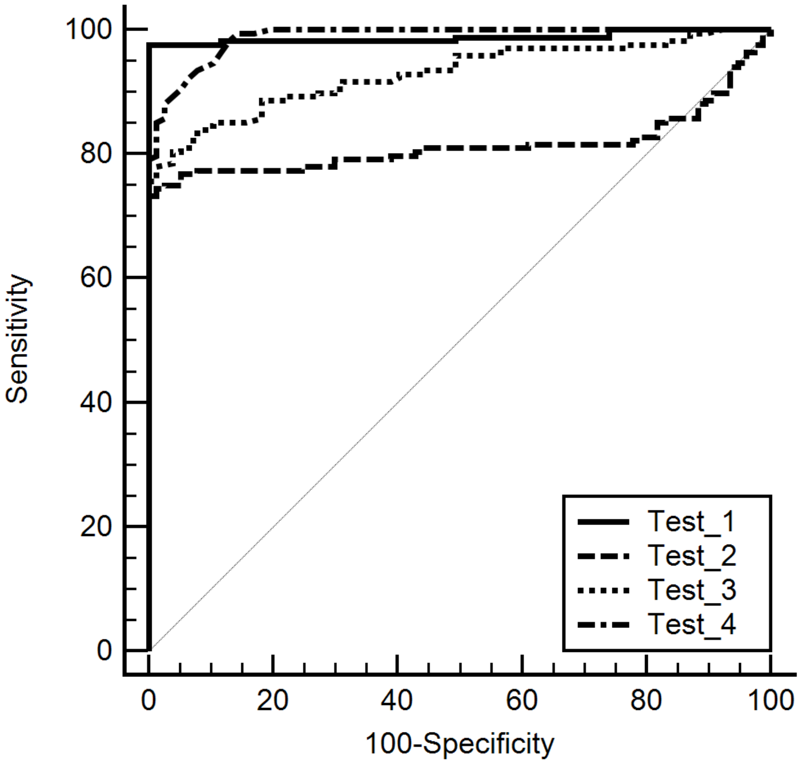


Table 5. Receiver operating characteristic (ROC) comparison of area under the curve (AUC) and p*-values for four different ELISAs to detect antibodies against influenza in pigs*

Parameter	Test 1	Test 2	Test 3	Test 4
Area under ROC curve	0.988	0.814	0.929	0.988

Table 6. Pairwise comparison of ROC curves

	Test 1 vs 4	Test 1 vs 3	Test 3 vs 4	Test 2 vs 3	Test 1 vs 2	Tetst 2 vs 4
Difference between AUC	0.0002	0.055	0.0589	0.115	0.173	0.174
95% CI	–0.016 to 0.016	0.025 to 0.092	0.027 to 0.090	0.069 to 0.161	0.117 to 0.230	0.118 to 0.229
Significance level	*p*=0.9808	*p* =0.0007	*p* =0.0003	*p* <0.0001	*p* <0.0001	*p* <0.0001

G. DISCUSSION AND CONCLUSIONS

Results from comparison experiments must be evaluated to reach a conclusion as to whether both methods are comparable using statistical analyses and objective assessments to assist with a final decision (NATA, 2013). However, often other criteria such as costs/equipment, throughput capacity, turn-around time, quality assurance capability, technical sophistication, acceptance in regulatory or scientific community and interpretative skills also need to be considered in decision-making (Figure 1, Step 6). It is important to have a process in place, which clarifies who has the authority to make the final decision whether methods are comparable, e.g. technical manager, laboratory director, quality manager.

For example, comparison of two TaqMan assays indicated strong and positive correlation. Repeatability of both methods was comparable when applying 2–3 SD or ± 2–3 Ct. Ct values from a weak positive-control sample were consistently lower for the M1 than for the N1 TaqMan, indicating a slightly superior Se of the M1 assay. If used as a screening test the M1 would be more suitable due to its higher Se. Results were corroborated by testing diagnostic samples from Hendra virus outbreaks between 2011 and 2013 (data not shown). On the other hand having two tests, which target different genes increases the chance of not missing a new variant. In the case of a deadly zoonosis such as Hendra virus this is an important consideration.

Results from the Bland–Altman Plot in Figure 4 and Table 4 show that statistical results sometimes are difficult to interpret, e.g. the 95% CI for the mean difference excludes zero. This might be interpreted as a lack of comparability but change in the cut-off value can be used to compensate for this outcome.

ROC analysis of 4 different swine influenza ELISAs (Figure 8 and Tables 5 and 6) indicated two tests of almost identical DSe and DSp (Tests 1 and 4). At the same time it helps to rank the performance of the other tests. Under these circumstances cost, availability and other criteria will determine the final decision as to which test is the most suitable for a designated purpose.

On the other hand, not all parameters need to be addressed in every method comparison study. For example, an assessment of cross-contamination would be necessary for changes in equipment such as manual vs robotic nucleic acid extraction procedure but not for changing a key reagent, e.g. a different primer or probe. It is good practice to decide on relevant parameters and acceptance/rejection criteria before the experiment. The most relevant question is whether the new test is fit-for-purpose.

It is in the nature of veterinary diagnostic testing to allow for flexibility when it comes to the specification of acceptance limits, e.g. when comparing two different molecular tests it could be that the difference is no more than 1, 2 or 3 Ct in 95% (99%) of tested samples, no more than 10% of the average of the 2 samples in at least 95% (99%) of tested samples, no more than 1, 2, or 3 SD of the samples. It is recommended to consider relevant parameters and limits up front although the qualitative considerations (e.g. cost, ease of testing, rapidity of results) probably come in at the end. Whatever parameter is chosen as a basic rule the candidate test should on average not be significantly underperforming the validated test.

The data generated and the decision making process for the acceptability of the change should be clearly documented and retained to show an audit trail.

H. DATA ANALYSIS

Data were stored and grouped in Microsoft Excel. Analysis and plotting for scatter diagrams, histograms, data distribution and plotting, Bland Altman Plots, comparison graphs and ROC analysis were performed using MedCalc (MedCalc®, Version 12.4.0.0, 64 bit, Window XP/Vista 7/8, www.medcalc.org, Copyright 1993–2013, MedCalc software bvba).

REFERENCES

BLAND J.M. & ALTMAN D.G. (1999). Measuring agreement in methods comparison studies. *Stat. Methods Med. Res.*, **8**, 135–160.

BLAND J.M. & ALTMAN D.G. (2007). Agreement between methods of measurement with multiple observations per individual. *J. Biopharm. Stat.*, **17**, 571–582.

BROCCHI E., BERGMANN I.E., DEKKER A. PATON D.J., SAMMIN D.J., GREINER M., GRAZIOLI S., DE SIMONE F., YADIN H., HAAS B., BULUT N., MALIRAT V., NEITZERT E., GORIS N., PARIDA S., SØRENSEN K. & DE CLERCQ K. (2006). Comparative evaluation of six ELISAs for the detection of antibodies to the non-structural proteins of foot-and-mouth disease virus. *Vaccine*, **24**, 6966–6979.

ENGEL B., BUIST W., ORSEL K., DEKKER A., DE CLERCQ C., GRAZIOLI S. & VAN ROERMUND H. (2008). A Bayesian evaluation of six diagnostic tests for food-and-mouth disease for vaccinated and non-vaccinated cattle. *Prev. Vet. Med.*, **86**, 124–138.

GALL D., COLLING A., MARINO O., MORENO E., NIELSEN K., PEREZ B. & SAMARTINO L. (1998). Enzyme immunoassays for serological diagnosis of bovine brucellosis: a trial in Latin America. *Clin. Diagn. Lab. Immunol.*, **5**, 654–651.

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

KOZAK M. & WNUK A. (2014). Including the Tukey mean-difference (Bland–Altman) plot in a statistics course. *Teaching Statistics*, **36**, 83–87.

NATIONAL ASSOCIATION OF TESTING AUTHORITIES (OF AUSTRALIA) (NATA) (2018). NATA General Accreditation Guidance – Validation and Verification of Quantitative and Qualitative Test Methods. https://www.nata.com.au/phocadownload/gen-accreditation-guidance/Validation-and-Verification-of-Quantitative-and-Qualitative-Test-Methods.pdf (accessed 22 November 2018)

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

*
* *

NB: FIRST ADOPTED IN 2016.