Factors affecting test reproducibility among laboratories

C. Waugh & G. Clark*

Australian Centre for Disease Preparedness, CSIRO, 5 Portarlington Road, 3219 East Geelong, Victoria, Australia

*Corresponding author: car577@csiro.au

Summary

Reproducibility is the ability of an assay to provide consistent results (when testing the same samples) in different laboratories. The validation of a new diagnostic assay should include specific assessment of assay reproducibility to determine the degree to which results are unaffected by minor changes in experimental conditions. Ideally, assessment of reproducibility involves the testing of identical samples in multiple laboratories by multiple analysts using the same method, reagents and controls, albeit with different equipment. Such an assessment will provide estimates of the precision and accuracy of an assay across laboratories. In reality, although the reproducibility of an assay is often assessed by multiple laboratories testing identical samples, the reagents, controls and testing platforms used, while similar, are usually not the same. Thus, reproducibility testing permits the assessment of variability resulting from different testing platforms, reagent supplies and operators. The determination of minor versus major variations in test conditions that may be anticipated in multi-laboratory use is part of the assessment at this stage of validation. Once validated, there are ongoing monitoring requirements to assess the performance characteristics and ensure they are consistently maintained. The use of quality assurance programmes is required, as this offers continued monitoring of assay performance by measuring the precision and accuracy of results for well-characterised samples and controls. Tests recommended by the World Organisation for Animal Health as fit for purpose are widely used internationally and need to have satisfactory reproducibility.

Keywords

Monitoring – Quality controls – Reproducibility – Validation – Variation.

Introduction

The repeatability and reproducibility of an assay are important measures of assay robustness (or ruggedness), that is, the capacity of an assay to remain unaffected by minor alterations to test protocols that may occur over time within a single laboratory or when deployed into many

https://doi.org/10.20506/rst.40.1.3213

laboratories (1). Crowther *et al.* (2) succinctly define ruggedness as 'a measure of resistance to user variables'.

Staff, technical training, reagents, equipment and laboratory infrastructure are variables that may impact the reliability of diagnostic results and are commonly evaluated by assessing assay repeatability and reproducibility.

When using well-characterised samples, repeatability and reproducibility provide data on the precision and accuracy of a test. Although the terms are often used interchangeably, repeatability and reproducibility define distinct concepts when applied to diagnostic assays. Repeatability measures the variation in test results obtained from testing identical samples in a single laboratory using a single operator (intra-assay repeatability) or testing identical samples in a single laboratory by two or more analysts (inter-assay repeatability). Implicit in repeatability testing is the use of a single protocol and identical reagents and controls. In contrast, reproducibility measures an assay's capacity to remain unaffected by changes or substitutions in test conditions when deployed across multiple laboratories and regions. It assesses the consistency of results (precision and accuracy) obtained from identical samples tested in different laboratories, how close the quantitative results are to each other and if this can be consistently repeated. During assay validation, reproducibility assessments are often (and ideally) conducted using reagents and controls distributed by the laboratory that has developed the test, but using the available equipment and testing platforms, which often vary across a network of laboratories. Once an assay has been satisfactorily validated, its full deployment to external laboratories may give rise to additional sources of variation as laboratories source reagents from a variety of suppliers, the number of analysts performing a test increases, and there are changes in testing platforms (such as thermocyclers, absorbance readers, etc.). Both the OIE and, in Australia, the National Association of Testing Authorities (the independent agency for accrediting laboratories) provide extensive advice on how to assess and minimise sources of variability that impact reproducibility. This paper will discuss the broad issues that may adversely impact assay variability and common means of monitoring assay reproducibility.

Assay development versus assay validation

During assay development, many factors will affect assay performance, and these must be assessed and optimised to produce a standardised protocol that is both sensitive and specific to the analyte being measured. Of necessity, assay development will involve extensive titration of reagents and assessment of critical components (primers, antibodies, buffer types and pH, etc.) as well as assay conditions (incubation times, shaker speeds, CO_2%, etc.) and result interpretation

(cut-off values and thresholds, definition of positive and negative results and result interpretation). The assessment and optimisation of test parameter performance is critical for determining whether or not an assay is provisionally accepted and whether or not it progresses through to validation. At the point the protocol is standardised, reagent concentrations and compositions are fixed within acceptable limits, controls are identified, and optimal test conditions and result interpretation are described, then validation can begin (OIE stage 1, see below). Thus, the validation of a candidate assay relies on the use of a clearly defined protocol with fixed limits to variables that may affect assay performance (3).

The OIE validation pathway can conveniently be divided into four stages that describe the characteristics to be assessed and the data required to impartially determine an assay's performance:

– Stage 1: analytical characteristics

– Stage 2: diagnostic characteristics

– Stage 3: reproducibility

– Stage 4: implementation (1).

Formal assessments of reproducibility are conducted in stages 1 and 3 of the pathway, and although the OIE offers provisional recognition of an assay that has completed stage 2, such recognition requires preliminary evidence of acceptable reproducibility. This can be provided somewhat by the intra-assay component of repeatability testing (different operators within the same laboratory), but it also requires small test panels of up to five samples to be issued to one or more highly competent laboratories where the test has been implemented and controlled (i.e. a scaled-down stage 3).

Acceptable repeatability is a pre-requisite of acceptable reproducibility, and it must be noted that acceptable repeatability within one laboratory is not sufficient for concluding that a test method exhibits acceptable reproducibility. By demonstrating that the test–retest variability of a sample under controlled conditions is constant, variability can be attributed only to errors caused by the measurement process itself (4). Reproducibility assessment aims to identify the variation in a test measurement under changing conditions.

Factors affecting test reproducibility among laboratories

Identifying and reducing controllable sources of variation in a test method is critical to retaining confidence in a test and its results (5) and in the quality of information that underpins decision-

making. Controllable sources of variation include the laboratory environment, technician capability, calibration of instruments, storage conditions, handling of reagents and samples, and quality of assay reagents. Distinguishing between technical variation and biological variation, and controlling these where possible, allows assessment of the measurement process. Most, if not all, of the controllable elements that may affect assay reproducibility are addressed by the process of laboratory accreditation to the international standard for testing and calibration laboratories, ISO/IEC 17025. The standard emphasises the central roles of staff competence and training, equipment calibration, quality management systems, use of validated assays (where available), verification of test performance when adopting or introducing new assays, use of internal quality control and external quality control, and participation in external proficiency testing (PT) programmes. All the laboratory's training, calibration and diagnostic processes must be documented and retained for audit purposes.

Even for smaller laboratories that may not have the resources necessary to implement accredited quality systems, the same principles can be used to improve laboratory performance.

It is beyond the scope of a single paper to describe the sources of variability that may affect reproducibility for the multitude of veterinary diagnostic assays that exist; however, it is possible to highlight themes common to all diagnostic laboratories that impact on test reproducibility. It is also useful to outline the major sources of variability that impact two major categories of veterinary diagnostic investigations, namely molecular and serological assays. These are summarised below.

Laboratory environment

In first-world laboratories with robust infrastructure, it is very much taken for granted that there will be a constant supply of electricity and that the temperature in the laboratory areas will be consistently moderate and uniform. In some remote or resource-limited areas, however, particularly in equatorial regions, these conditions cannot be guaranteed, and the ambient temperature of a non-air-conditioned (or intermittently air-conditioned) laboratory can increase the burden of operation on standard laboratory equipment such as fridges and freezers, reducing their effectiveness and lifespan. Poor storage of reagents due to inadequate refrigeration may cause significant deterioration of reagents (2) to the point that assays fail quality assurance (QA) assessments. Poor storage often causes decreased shelf-life that may contribute to assay variability and confound the investigation of failed QA measures in assay runs or PT assessments. Some mitigation of these effects may be achieved by aliquoting reagents into single use (or limited use) vials, maintaining concentrated stocks of critical components and developing

robust house-keeping protocols to track the use of reagents and the storage conditions (temperature monitoring of fridges/freezers). These measures are often routine in well-resourced laboratories, where storage space and access to consumables (such as vials and storage systems, racks, boxes, etc.) is relatively unrestricted and where robust procedures are in place (e.g. digital temperature probes and automated temperature surveillance). It is hard to overstate the importance of minimising freeze/thaw cycles on reagents to maintain assay consistency and accuracy.

During seasonal weather disturbances, such as monsoons, power outages can be commonplace, adding to the demand on laboratory equipment (6). Power outages may have the immediate effect of disturbing assays in progress, but they also often have a cumulative deleterious effect on laboratory equipment reliability. The use of uninterruptable power supply (UPS) units on critical equipment is highly recommended in order to maintain equipment performance; however, these require a significant financial investment.

The maintenance of stable laboratory environmental conditions (air temperature, air handling, gas and electricity) and regular maintenance of equipment are mandated in international standards relating to diagnostic laboratories, in recognition of the importance of environment and equipment to the generation of accurate and precise results. These standards include pertinent laboratory standards (such as ISO 17025 for veterinary diagnostics or ISO 15189 for medical laboratories) but also ISO 9001, which more broadly addresses an organisation's quality management systems, their internal auditing processes (to identify and address risks) and their ongoing focus on improvement. Regular maintenance schedules impose a significant financial burden and, in some geographical areas, they are complicated by a lack of availability of trained technicians. They are, however, essential to ensuring reliable equipment and robust assay performance.

Technician capability

Standardisation of procedures, technical training, mentoring of junior staff and on-going professional development are essential to developing capable, knowledgeable staff that are both competent in assay performance and able to assess QA data and recognise issues or trends in data as they arise. QA data is accumulated every time a control sample is analysed, typically alongside diagnostic specimens, and can be used to monitor the performance of the assay and that of the operator. To maximise the utility of QA data it is important to prepare large, homogeneous, stable batches of controls that can be used longitudinally to monitor equipment, assays and operator performance. It is also important to review data on a regular basis to detect

shifts in statistical parameters or any clustering of results that require investigation. These changes may indicate, for example, that there are gaps in training, that there has been some deterioration of controls, reagents or equipment, or that equipment needs to be serviced and/or recalibrated.

Well-structured, comprehensive training programmes are required to standardise the training of new staff and provide refresher training for established staff or staff returning from extended periods of leave. During training, the use of known positive and negative controls, provided 'blind' to an operator, is important for demonstrating competence in an assay (rather like an internal PT assessment). Likewise, titration of strong controls can be used to assess assay sensitivity and may provide insights into an operator's accuracy and pipetting skill and their ability to repeat the test in a consistent manner.

Standardised methods, standardised staff training, identification of on-going training opportunities for staff, and mentoring of new staff by knowledgeable and competent technicians all contribute to the development of a work practice that enhances overall laboratory competence and capability.

Calibration of instruments

Pipettes are often the workhorse of the laboratory and even the most skilled technician is only as good as the equipment used. Pipettes must be regularly assessed for accurate and repeatable functioning, especially at the extremes of their range, and accurate records kept of calibration assessments.

As noted above, critical equipment must undergo regular servicing and calibration to ensure optimal functioning. It is good practice to record all equipment used for an assay run and to standardise the use of critical components to maximise the repeatability of intra- and inter-assay runs.

Consistency and quality of assay reagents

Changes in suppliers of reagents, changes in the formulation of established products or even batch-to-batch variation of reagents can impact assay repeatability and/or reproducibility. In well-resourced laboratories, reagents are chosen primarily on the basis of the quality of reagents and independent assessments of fitness-for-purpose; however, in some regions, financial constraints may determine which reagents or kits are purchased. The consistency and quality of reagents or kits can have a significant impact on assay reproducibility. Verification studies (see

Kirkland and Newberry in this issue [7]) are necessary to ensure that changes in suppliers of reagents, changes in reagent formulations, or changes in batches of critical assay components are acceptable and that assays are (and remain) fit for purpose. It is useful to assess the consistency and quality of reagents using well-characterised controls, or reference material with known performance criteria, and to titrate these to extinction (that is, titrate to below the level of detection) to assess any effects on sensitivity. Measures repeated by two or more operators can provide data on assay repeatability, accuracy and precision.

As noted previously, appropriate storage of reagents is essential to maximise consistency and performance. Aliquoting reagents into small, concentrated volumes can limit freeze/thaw cycles and/or minimise the risk of introducing contamination into reagent stocks. If contamination is detected or suspected, new aliquots of stocks can be used to assist with trouble-shooting and determining sources of contamination.

Factors affecting reproducibility of molecular assays

In the veterinary diagnostic setting, molecular assays are developed usually for simple agent detection, providing a qualitative diagnosis (positive or negative), rather than a quantitative estimate of agent genome copy number or virion/microbe particles. Some research and clinical (mostly human) molecular assays may provide quantitative estimates, either relative or absolute, of gene expression, providing evidence of disease progression or viral load over time. There are significant issues relating to the standardisation of quantification of gene expression that affect the reproducibility of such molecular assays and there is a growing literature on how these may be addressed (8, 9). Guidelines have been published to assist with standardising the language and methodology of quantitative real-time polymerase chain reaction (PCR) experiments. They include thorough explanations of areas that may impact reproducibility and the minimum information that is required for publication so that methods, results and interpretations can be appropriately scrutinised (10). Many of these areas are also important to the reproducibility of qualitative assays, even when gene expression quantification, as an absolute or relative measure, is not performed.

For molecular assays, protocol details that commonly differ between laboratories and may impact assay reproducibility (and sensitivity) include choice of nucleic acid extraction method (11), choice of enzyme and master mix, amount of template used in PCR, introduction of redundant nucleotides into primer/probe sequences and, for real-time PCR, choice of DNA dyes and probe chemistries. The repeatability and reproducibility of real-time molecular assays can be affected by technical issues such as pipette calibration and even pipetting technique, especially

regarding the addition of template, which usually comprises 10% or less of the reaction volume. As a doubling of DNA template reduces a cycle threshold (Ct) value by one unit, replicate analyses can provide information on assay repeatability. Replicate Ct values that are within 0.5 units of each other usually indicate good pipetting technique, well-calibrated pipettes and stable, reliable reagents.

The reproducibility of molecular assays, particularly real-time assays, can also be affected by PCR inhibitors and contamination of reagents (12, 13). Uni-directional work flows, clean rooms for preparation of master mix and storage of reagents away from sources of end-product DNA can all contribute to optimal assay performance.

It is of paramount importance that any change to a method undergoes verification, that is, side-by-side comparison between established and new protocols or reagents to assess the effect of changes on assay accuracy and precision. This also extends to new batches of established reagents (sourced either in-house or commercially) to ensure batch-to-batch consistency of primers, master mix, enzymes, etc. The titration of well-characterised, strong-positive controls using current and new batches of reagents or current and new protocols provides a ready assessment of changes in assay sensitivity that may impact reproducibility of results. Verification studies should also include repeated measures over time using two or more operators to gauge assay repeatability, accuracy and precision. More broadly, guidelines for the validation of real-time PCR assays for veterinary diagnostics, including aspects of repeatability and reproducibility, have been developed by the Laboratory Technology Committee of the American Association of Veterinary Laboratory Diagnosticians (14). The guidelines include detailed examples of method comparability studies for the conversion of singleplex assays into multiplex assays and include a comprehensive discussion of best practice for the performance of verification studies.

Factors affecting reproducibility of serological assays

Factors affecting reproducibility of serological assays include the quality and quantity (titre) of antibodies (primary and secondary antibody), the class of antibody and the epitope recognised, which can affect the avidity of reactions. Blocking reagents, water and buffer quality, water hardness, and the composition of plastic plates or other membrane supports (e.g. nitrocellulose strips) can impact protein binding, assay sensitivity and assay repeatability and reproducibility. In addition, different plate readers within a laboratory may give different absorbance measurements due to the relative age of the light source. Consequently, serological assays often

display a greater spread of results compared to molecular assays and, as noted previously, best practice recommends the standardisation of equipment used for critical functions.

Comprehensive assessments of factors that impact the reproducibility of enzyme-linked immunosorbent assays (ELISA) are provided in two papers produced from the joint initiative of the Food and Agriculture Organization of the United Nations and the International Atomic Energy Agency on the diagnosis and surveillance of livestock diseases (2, 6). The agencies have worked extensively in under-resourced countries to help establish standardised procedures for the use of ELISA kits for various animal and epizootic pathogens. The papers emphasise the importance of all factors discussed herein on the reproducibility of results and additionally discuss the challenges of standardising documentation and test interpretation. They also discuss the roles of the kit producer and end-users in maximising the performance of ELISA kits.

As for any change in protocols or reagents, verification studies must be performed using well-characterised controls in side-by-side comparisons. Studies must include statistical analysis of results in order to provide qualitative and quantitative measures of intra- and inter-assay repeatability (precision). Where it is possible to test controls with a known concentration of analyte, repeated measures will provide assessment of accuracy and precision.

Common methods of assessing reproducibility

Evaluation panels

As part of early test design, the intended purpose (the end use) of the assay is defined and, subsequently, that purpose (e.g. disease diagnosis, confirmatory assessment, surveillance etc.) should be taken into account when undertaking and preparing reproducibility assessments (15).

The OIE validation pathway for a new test recommends that reproducibility be assessed by having three or more laboratories perform the test on the same panel of samples in an inter-laboratory test comparison. The composition of the evaluation panel should reflect the inherent biological variability of the target pathogen (different serotypes) and circulating or geographically relevant strains. Where possible, the samples should be prepared from known positive field samples (e.g. leptospirosis-positive bovine urine) or from positive samples spiked into suitable matrix to mimic a true test sample (e.g. negative bovine urine spiked with a known concentration of leptospirosis bacteria). The panel should contain a full representation of concentrations covering the operating range of the assay in animals of the target population, to ensure evidence of reproducibility at a range of possible analyte levels, not just at either end of the scale, i.e. strong-positive and weak-positive.

Any samples used for reproducibility studies must be homogeneous and stable for the period of use, whether that be a short period of time, as in the case of a reproducibility study, or an extended period of time, as in the case of an external quality control programme. Where possible, large test batches should be prepared to enable testing over multiple testing intervals. For this purpose, it is useful to use samples that are well characterised and tested under repeatability conditions prior to selection. A PT provider accredited under the ISO/IEC 17043 standard can provide this level of sample assessment. For many test developers this may not be an option, but it is possible to develop procedures and documents to demonstrate that the samples are homogeneous and stable. Ongoing management and replacement of sample stocks is critical to ensure long-term access for ongoing validation requirements and to ensure that samples are available when method changes occur and when verification assessments are undertaken. A later paper within this issue discusses the use of Biobanks as a repository for ongoing access to and management of test panels and test samples for validation purposes (see Watson *et al.*, this issue [16]).

The OIE recommends that at least three laboratories should test the same panel of samples (blinded) using the same protocol, reagents and controls. The panel should contain 20 or more samples, with approximately 25% of the samples being negative and the remainder positive (17). All samples should be tested in duplicate, and some samples should be present in the panel two or more times to allow within-laboratory repeatability estimates. Table I provides an example of nine independent samples replicated to comprise a testing panel of 30 samples. Evaluation panels can also be represented in different formats, as implementation panels or inter-laboratory test panels, at different stages of the validation pathway. Reproducibility assessments are not only for the initial phase of validation but can also support ongoing monitoring requirements as part of a network approach. An example is demonstrated in the case study shown in Box 1, with the use of inter-laboratory test comparison.

Proficiency testing

Proficiency testing (PT) is an important component of ongoing assay reproducibility assessments. Although primarily designed to assess the competence of a laboratory in performing a diagnostic assay, PT can also provide an assessment of the reproducibility of a standardised test protocol that has been deployed across a laboratory network, if the network laboratories participate in the same PT programme. In Australia, the principal veterinary diagnostic laboratories in each state and territory jurisdiction comprise a network known as the laboratories for emerging animal disease diagnosis and response (LEADDR) network. In addition to any preferred international PT programmes, each jurisdictional laboratory also

participates in accredited PT programmes provided by the Australian Centre for Disease Preparedness (ACDP, formerly known as AAHL). The ACDP-PT programme focuses on tests that have been standardised and deployed across the LEADDR network, targeting endemic and exotic diseases of relevance across Australia. It is noteworthy that the PT programme for the LEADDR network is an open programme, with results revealed to all LEADDR participants. The programme provides multiple benefits (e.g. evidence of laboratory proficiency in diagnostic assessment and ongoing evidence of assay reproducibility) and it allows participants to discuss assay modifications that may be introduced by an individual laboratory. It also enables participants to assess the relative effect of these modifications as part of the verification process.

Even where results are coded and confidential, PT can provide proxy evidence that an assay is acceptably reproducible if the methods used by participants are known and tabulated so that results from similar methods can be compared. As the role of the PT provider is to assess the competence of laboratories for diagnostic testing, it is not usual practice for PT reports to include formal method comparisons. The case study (Box 1) included in this paper demonstrates how simple statistical approaches and knowledge of the test protocols used in inter-laboratory comparisons and PT can be used to identify clustering of results that may suggest there are variations in test protocols that affect reproducibility and assay sensitivity. Although participation and acceptable performance in accredited PT programmes is a necessary component of accreditation under ISO/IEC 17025, PT provides only a snapshot of laboratory competence and assay reproducibility. For ongoing assessment of reproducibility, alternative methods are required that provide reliable longitudinal assessments of assay precision and accuracy.

Internal and external quality controls

To maintain an assay's validation status, continued and routine monitoring of performance is required to ensure consistency of performance characteristics and monitor for systematic error (bias) (1). This may be done as an internal process, using internal quality controls (IQCs), and/or as an external process independent of the laboratory. For the purposes of this paper, IQCs are defined as any in-house controls prepared within a laboratory or a control that is provided as part of a kit and used to assess assay performance. External quality controls (EQCs) are defined as controls that are sourced and prepared independently of the laboratory or kit producer and these will be discussed in more detail below.

At a minimum, IQCs (positive and negative) provide assurance that a test assay has worked adequately. When prepared as large, homogeneous, stable batches, IQCs can also provide

ongoing assurance of assay fidelity and operator competence, and measures of intra- and inter-assay repeatability.

Assessment of reproducibility requires large batches of homogeneous, stable controls that can be distributed to laboratories using similar assays. Independent and externally-prepared material to monitor assay performance and reproducibility can be provided through participation in formal PT (as discussed above), through evaluation panels used in inter-laboratory test comparisons (such as those used in the assay validation pathway), or through the use of EQCs, such as international reference standards (IRS) and network quality controls (NQCs). IRS and NQCs are typically prepared from well-characterised field isolates with regional relevance, prepared in bulk and aliquoted, with documented homogeneity and stability data. EQCs may have a pre-assigned analyte concentration range and be used as part of a laboratory's quality assurance programme or, if results are returned to a central collation agency, can provide direct comparisons of inter-laboratory reproducibility based on either pre-assigned values or group medians. In contrast to PT, the use of EQCs in combination with IQCs provides longitudinal data on assay performance, including repeatability within individual laboratories, and assay reproducibility when laboratories use similar protocols. When distributed within a network of laboratories EQCs/NQCs can provide robust evidence of assay reproducibility and provide early detection of problems such as: adverse assay modifications, inferior reagents (from diverse suppliers or when supply chains change), and changes in commercial reagent formulations that affect assay performance. To maximise the sensitivity of EQCs/NQCs as monitoring tools, batches of controls should approach the limits of detection of the assay, as this is the region in which all assays display the greatest variability (least precision) and thus are most sensitive to adverse changes in assay protocols, reagents and testing platforms, all of which can impact reproducibility. Figure 1 is an example of an NQC distributed to a laboratory network, where each laboratory uses two or more real-time PCR methods for detection of classical swine fever (CSF), which, in this region (Australia), is an exotic agent. The methods used regularly were those published by Haines *et al.* (18), Hoffmann *et al.* (19), Eberling *et al.* (20) and Risatti *et al.* (21).

In this setting, the use of a quality-assured NQC, prepared at low viral titre, has several benefits:

– it provides assurance that each jurisdictional laboratory can detect the target agent, and that the assays in use are suitable for surveillance purposes

– it allows for the relative sensitivity of the assays for detecting CSF to be assessed by a comparison of the mean Ct value and the spread of results for each assay

– it provides a quantitative measure of the reproducibility of each assay by illustrating the spread of the data with repeated measures over time.

Each laboratory, represented by a different colour in the figure, can monitor their performance against the mean and standard deviation and in relation to the performance of other laboratories within the network for each of the assays.

IQCs and EQCs are used to monitor test performance and detect departures from set norms via pre-determined quality indicators (or limits) and predefined acceptance criteria in order to detect random and systematic analytical errors. Monitoring quality control processes through predetermined quality indicators identifies the distribution of measured outputs, using statistical methods to determine assignable cause variation in the presence of background noise, thus helping to classify the process as in, or out, of control (13).

Descriptive statistics and control charts are often used to monitor dispersion of and shifts in quality control results through mean, standard deviations, variance, probability distribution factors, range etc., with the addition of multirules (22) (e.g. Westgard rules) alongside a Levey-Jennings chart (23, 24) to set limits (see Table II and Fig. 2). Clear rules should define limits and actions to be taken to investigate and correct unacceptable variance.

Conclusions

Reproducibility testing is a central element of assay validation when it is used to assess the robustness of a protocol undergoing technology transfer and deployment to external laboratories. Reproducibility assessment is equally important to the monitoring of established tests to provide evidence of suitable assay, equipment and operator performance. Implicit to the acquisition of reproducibility data is the preparation in bulk of stable, homogenous, relevant material that can be tested by diverse laboratories using the same test method. Although, ideally, laboratories should be using identical protocols and reagents, in reality, this is seldom the case, especially where laboratories are widely dispersed and use diverse suppliers and equipment platforms. The testing of identical samples by laboratories using *similar* protocols thus becomes a measure of assay robustness and a means of identifying the limits of variations in test conditions that impact test reproducibility, precision and accuracy.

Les facteurs affectant la reproductibilité des tests dans différents laboratoires

C. Waugh & G. Clark

Résumé

La reproductibilité d'un test désigne son aptitude à fournir des résultats constants (en testant les mêmes échantillons) lors d'analyses effectuées par différents laboratoires. Toute validation d'un nouvel essai diagnostique doit inclure une étape spécifique d'évaluation de la reproductibilité de l'essai, visant à vérifier jusqu'à quel point les résultats du test demeurent inchangés en cas de légères variations dans les conditions de réalisation de l'essai. Idéalement, l'évaluation de la reproductibilité consiste à soumettre des échantillons identiques à une même méthode d'essai réalisée dans plusieurs laboratoires par des analystes à chaque fois différents et en utilisant les mêmes réactifs et contrôles, mais avec des équipements différents. Une telle évaluation fournit une estimation de la précision et de l'exactitude d'un essai conduit par plusieurs laboratoires. Dans la pratique, si la reproductibilité d'un essai est souvent évaluée par des laboratoires différents à partir d'échantillons similaires, les réactifs, les contrôles et les plateformes d'essai ne sont généralement pas les mêmes d'un laboratoire à l'autre, bien qu'étant similaires. Les essais de reproductibilité permettent ainsi d'évaluer la variabilité induite par l'utilisation de plateformes de test et de réactifs différents par des opérateurs eux-mêmes différents. La détermination du caractère mineur ou substantiel des variations des conditions d'essai susceptibles d'être anticipées dans un cadre d'utilisation impliquant de nombreux laboratoires fait partie de l'évaluation à ce stade de la validation. Une fois la validation faite, des exigences de contrôle continu sont requises afin d'évaluer les caractéristiques de performances et de veiller à leur maintien dans le temps. Les laboratoires doivent se doter d'un programme d'assurance qualité qui garantisse le suivi continu des performances des essais à travers une mesure de la précision et de l'exactitude des résultats à partir d'échantillons et de contrôles bien caractérisés. Les tests dont l'aptitude à l'emploi selon l'objectif prévu a été établie et qui de ce fait sont recommandés par l'Organisation mondiale de la santé animale ont vocation à être utilisés dans le monde entier ; il est donc important qu'ils soient suffisamment reproductibles.

Mots-clés

Contrôle de la qualité – Reproductibilité – Suivi – Validation – Variation.

Factores que inciden en la reproducibilidad de las pruebas entre laboratorios

Resumen

Se entiende por “reproducibilidad” el conjunto de características que permiten que un ensayo depare resultados uniformes al ser aplicado a las mismas muestras en laboratorios distintos. El proceso de validación de una prueba de diagnóstico debe incluir una evaluación específica de su reproducibilidad, a fin de determinar en qué medida los resultados se mantienen inalterados ante cambios menores de las condiciones experimentales. Lo idóneo para evaluar la reproducibilidad es que, en múltiples laboratorios, múltiples analistas sometan a prueba muestras idénticas empleando idénticos métodos, reactivos y controles, pero distinto equipo de laboratorio. Semejante evaluación permitirá estimar la precisión y exactitud que ofrece un ensayo en diferentes laboratorios. En realidad, aunque en la evaluación de la reproducibilidad de un ensayo intervienen a menudo múltiples laboratorios que analizan muestras idénticas, los reactivos, controles y dispositivos de prueba utilizados, aún siendo parecidos, no suelen los mismos. La realización de pruebas de reproducibilidad permite pues determinar la variabilidad que introducen distintos dispositivos de prueba, suministros de reactivos y técnicos de laboratorio. Así, la determinación de las variaciones menores frente a las variaciones importantes de las condiciones experimentales que cabe prever cuando múltiples laboratorios emplean un ensayo forma parte de la evaluación en esta fase del proceso de validación. Una vez validado un ensayo, hay una serie de requisitos de seguimiento continuo que sirven para evaluar las características de rendimiento y garantizar que se mantengan estables en el tiempo. Para ello es necesario utilizar programas de garantía de calidad, que ofrecen la posibilidad de hacer un seguimiento continuo del rendimiento de un ensayo cuantificando la precisión y exactitud de los resultados obtenidos con muestras y controles bien caracterizados. Las pruebas que la Organización Mundial de Sanidad Animal recomienda por considerarlas adaptadas a su finalidad, de uso muy extendido a escala internacional, deben presentar un nivel satisfactorio de reproducibilidad.

Palabras clave

Controles de calidad – Reproducibilidad – Seguimiento – Validación – Variación.

References

1. National Association of Testing Authorities (NATA) (Australia) (2018). – General accreditation guidance. Validation and verification of quantitative and qualitative test methods. NATA, Sydney, Australia, 31 pp. Available at: www.nata.com.au/phocadownload/gen-accreditation-guidance/Validation-and-Verification-of-Quantitative-and-Qualitative-Test-Methods.pdf (accessed on 27 January 2021).

2. Crowther J.R., Unger H. & Viljoen G.J. (2006). – Aspects of kit validation for tests used for the diagnosis and surveillance of livestock diseases: producer and end-user responsibilities. *Rev. Sci. Tech. Int. Off. Epiz.,* **25** (3), 913–935. doi:10.20506/rst.25.3.1706.

3. World Organisation for Animal Health (OIE) (2019). – Chapter 1.1.6. Principles and methods of validation of diagnostic assays for infectious diseases. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 28th Ed. OIE, Paris, France. Available at: www.oie.int/standard-setting/terrestrial-manual/access-online/ (accessed on 27 January 2021).

4. Bartlett J.W. & Frost C. (2008). – Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet. Gynecol.,* **31** (4), 466–475. doi:10.1002/uog.5256.

5. Dargatz D.A., Byrum B.A., Collins M.T., Goyal S.M., Hietala S.K., Jacobson R.H., Kopral C.A., Martin B.M., McCluskey B.J. & Tewari D. (2004). – A multilaboratory evaluation of a commercial enzyme-linked immunosorbent assay test for the detection of antibodies against *Mycobacterium avium* subsp. *paratuberculosis* in cattle. *J. Vet. Diagn. Invest.,* **16** (6), 509–514. doi:10.1177/104063870401600604.

6. Jeggo M.H. (2000). – An international approach to laboratory diagnosis of animal diseases. *Ann. NY Acad. Sci.,* **916** (1), 213–221. doi:10.1111/j.1749-6632.2000.tb05292.x.

7. Kirkland P.D. & Newberry K.M. (2021). – Your assay has changed – is it still 'fit for purpose'? What evaluation is required? *In* Diagnostic test validation science: a key element for effective detection and control of infectious animal diseases (I.A. Gardner & A. Colling, eds). *Rev. Sci. Tech. Off. Int. Epiz.*, **40** (1), XX–YY. doi:10.20506/rst.40.1.XXXX.

8. Derveaux S., Vandesompele J. & Hellemans J. (2010). – How to do successful gene expression analysis using real-time PCR. *Methods*, **50** (4), 227–230. doi:10.1016/j.ymeth.2009.11.001.

9. Bustin S. & Nolan T. (2017). – Talking the talk, but not walking the walk: RT-qPCR as a paradigm for the lack of reproducibility in molecular research. *Eur. J. Clin. Invest.*, **47** (10), 756–774. doi: 10.1111/eci.12801.

10. Bustin S.A., Benes V., Garson J.A., Hellemans J., Huggett J., Kubista M., Mueller R., Nolan T., Pfaffl M.W., Shipley G.L., Vandesompele J. & Wittwer C.T. (2009). – The MIQE

guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.*, **55** (4), 611–622. doi:10.1373/clinchem.2008.112797.

11. Fleige S. & Pfaffl M.W. (2006). – RNA integrity and the effect on the real-time qRT-PCR performance. *Mol. Aspects Med.,* **27** (2–3), 126–139. doi:10.1016/j.mam.2005.12.003.

12. Schrader C., Schielke A., Ellerbroek L. & Johne R. (2012). – PCR inhibitors: occurrence, properties and removal. *J. Appl. Microbiol.,* **113** (5), 1014–1026. doi:10.1111/j.1365-2672.2012.05384.x.

13. Lorenz T.C. (2012). – Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *J. Vis. Exp.,* **63**, e3998. doi:10.3791/3998.

14. Toohey-Kurth K., Reising M.M. […] & Crossley B.M. (2020). – Suggested guidelines for validation of real-time PCR assays in veterinary diagnostic laboratories. *J. Vet. Diagn. Invest.,* **32** (6), 802–814. doi:10.1177/1040638720960829.

15. Burd E.M. (2010). – Validation of laboratory-developed molecular assays for infectious diseases. *Clin. Microbiol. Rev.,* **23** (3), 550–576. doi:10.1128/cmr.00074-09.

16. Watson J.W., Carlile G.A. & Williams D.T. (2021). – The value of virtual biobanks for transparency purposes with respect to reagents and samples used during test development and validation. *In* Diagnostic test validation science: a key element for effective detection and control of infectious animal diseases (I.A. Gardner & A. Colling, eds). *Rev. Sci. Tech. Off. Int. Epiz.*, **40** (1), XX–YY. doi:10.20506/rst.40.1.XXXX.

17. World Organisation for Animal Health (OIE) (2019). –. Chapter 2.2.6. Selection and use of reference samples and panels. *In* Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, 28th Ed. OIE, Paris, France. Available at: www.oie.int/standard-setting/terrestrial-manual/access-online/ (accessed on 27 January 2021).

18. Haines F.J., Hofmann M.A., King D.P., Drew T.W. & Crooke H.R. (2013). – Development and validation of a multiplex, real-time RT PCR assay for the simultaneous detection of classical and African swine fever viruses. *PloS One,* **8** (7), e71019. doi:10.1371/journal.pone.0071019.

19. Hoffmann B., Beer M., Schelp C., Schirrmeier H. & Depner K. (2005). – Validation of a real-time RT-PCR assay for sensitive and specific detection of classical swine fever. *J. Virol., Methods,* **130** (1–2), 36–44. doi:10.1016/j.jviromet.2005.05.030.

20. Eberling A.J., Bieker-Stefanelli J., Reising M.M., Siev D., Martin B.M., McIntosh M.T. & Beckham T.R. (2011). – Development, optimization, and validation of a classical swine fever virus real-time reverse transcription polymerase chain reaction assay. *J. Vet. Diagn. Invest.,* **23** (5), 994–998. doi:10.1177/1040638711416970.

21. Risatti G.R., Callahan J.D., Nelson W.M. & Borca M.V. (2003). – Rapid detection of classical swine fever virus by a portable real-time reverse transcriptase PCR assay. *J. Clin. Microbiol.,* **41** (1), 500–505. doi:10.1128/jcm.41.1.500-505.2003.

22. Walker B.S., Pearson L.N. & Schmidt R.L. (2020). – An analysis of multirules for monitoring assay quality control. *Lab. Med.,* **51** (1), 94–98. doi:10.1093/labmed/lmz038.

23. Westgard J., Groth T., Aronsson T., Falk H. & Verdier C. (1977). – Performance characteristics of rules for internal quality control: probabilities for false rejection and error detection. *Clin. Chem.,* **23** (10), 1857–1867. doi:10.1093/clinchem/23.10.1857.

24. Westgard J.O. & Groth T. (1979). – Power functions for statistical control rules. *Clin. Chem.,* **25** (6), 863–869. doi:10.1093/clinchem/25.6.863.

Box 1

Prior to the commencement of surveillance activities, five laboratories across different jurisdictions were assessed to co at these laboratories were reproducible and conformed with the expected results for well-characterised samples. Five la laboratory test comparison and were provided with a blinded evaluation panel of 30 samples, comprising 8 independent to comprise a testing panel of 20 positive and 10 negative samples. The positive samples (strong, moderate and weak), assessment and were repeated within the panel as sample pairs to assess between-laboratory (inter-assay) reproducibi laboratory (intra-assay) repeatability estimates.

Data for a strong, moderate and weak sample set are presented in Figure 3 using Youden plots to illustrate the spread c shows the results for a sample pair. For each laboratory, the result for one sample is mapped on the x-axis and the resu y-axis. The Youden plot gives an immediate idea of the dominating sources of error in the results. Across each plot the e right, indicating that the data is dominated by systematic variation.

Testing was undertaken under the quality framework for ISO/IEC 17025, with each laboratory accredited to this standard quantitative real-time PCR (qPCR) method was used by each laboratory; however, several method variables were not c sources of method variability were: the detection platform in use (3 variables); the extraction method (2 variables); probe primer and probe concentrations (2 variables); master mix reagents (2 variables); template volume (2 variables); and se clustering of laboratory groups can be observed between each Youden plot for each sample set (Fig. 3 a, b, c). This cor between samples indicates a consistent influence affecting the reproducibility of the test method. The sources of variabi systematic influence on the data, i.e. results for Lab 1 (dark-blue triangle) consistently clustered outside the ellipse in the (indicating better sensitivity in relation to the group median), and Lab 5 (square) was consistently logged in the upper-rig relation to the group median). Despite the observed clustering of results, the coefficient of variation (CV) remained at 6% CV was lowest in the weak-positive sample, where an assay is most likely to display the greatest variability (least precis variables (probe chemistry and primer and probe concentrations) that were not consistent with any other laboratory meth method variables it is not possible to determine which of these variables or collection of variables is the cause of the obs reporting laboratories as a collective. The impact of systematic differences between laboratories does need to be consid purpose and application of the assay.

Table I

Example of blinded test panel composition of 30 samples, with 10 negative and 20 positive samples representing a range of test values for real-time polymerase chain reaction

The panel consists of nine independent samples replicated within the panel to provide estimates for both between-laboratory and within-laboratory reproducibility and repeatability estimates

Sample	Test samp	Expect
1	18-02 Positive - 1/3 dilution	23.7
2	18-03 Positive - 1/10 dilution	25.3
3	18-09 Negative	negative
4	18-06 Positive - 1/300 dilution	30.4
5	18-04 Positive - 1/30 dilution	25.9
6	18-09 Negative	negative
7	18-08 Positive - 1/4000 dilution	34.5
8	18-02 Positive - 1/3 dilution	23.7
9	18-09 Negative	negative
10	18-07 Positive - 1/300 dilution	32.3
11	18-09 Negative	negative
12	18-05 Positive - 1/100 dilution	25.6
13	18-09 Negative	negative
14	18-06 Positive - 1/300 dilution	30.4
15	18-09 Negative	negative
16	18-01 Positive - undiluted	22.4
17	18-07 Positive - 1/300 dilution	32.3
18	18-09 Negative	negative
19	18-03 Positive - 1/10 dilution	25.3
20	18-09 Negative	negative
21	18-03 Positive - 1/10 dilution	25.3

22	18-07 Positive - 1/300 dilution	32.3
23	18-04 Positive - 1/30 dilution	25.9
24	18-01 Positive - undiluted	22.4
25	18-09 Negative	negative
26	18-09 Negative	negative
27	18-05 Positive - 1/100 dilution	25.6
28	18-08 Positive - 1/4000 dilution	34.5
29	18-05 Positive - 1/100 dilution	25.6
30	18-08 Positive - 1/4000 dilution	34.5

Table II

Westgard rules are often established alongside a Levey-Jennings chart (Fig. 2) to set control limits (8)

Rule	Definition
1_{2S}	Warning when one control observation exceeds the mean by +/- 2 standard deviations (SD). Note: The applied once this warning has been given
1_{3S}	Reject when one observation exceeds the mean by +/- 3 SD (random error)
2_{2S}	Reject when two consecutive observations exceed the mean by either +2 standards or -2 SD (systemati
R_{4S}	Reject when one observation exceeds the mean by +2 SD and the next exceeds it by -2 SD (random er
4_{1S}	Reject when four consecutive observations exceed the mean by either +1 SD or -1 SD (systematic error
10x	Reject when ten consecutive observations fall on one side of the mean (systematic error)

1_{2S} – one value exceeding 2 SD
1_{3S} – one value exceeding 3 SD
2_{2S} – two consecutive results exceeding 2 SD
R_{4S} – two consecutive results with one greater than 2 SD and one less than 2 SD
4_{1S} – four consecutive results exceeding 1 SD (all in the same direction)
10x – ten consecutive results on one side of the mean

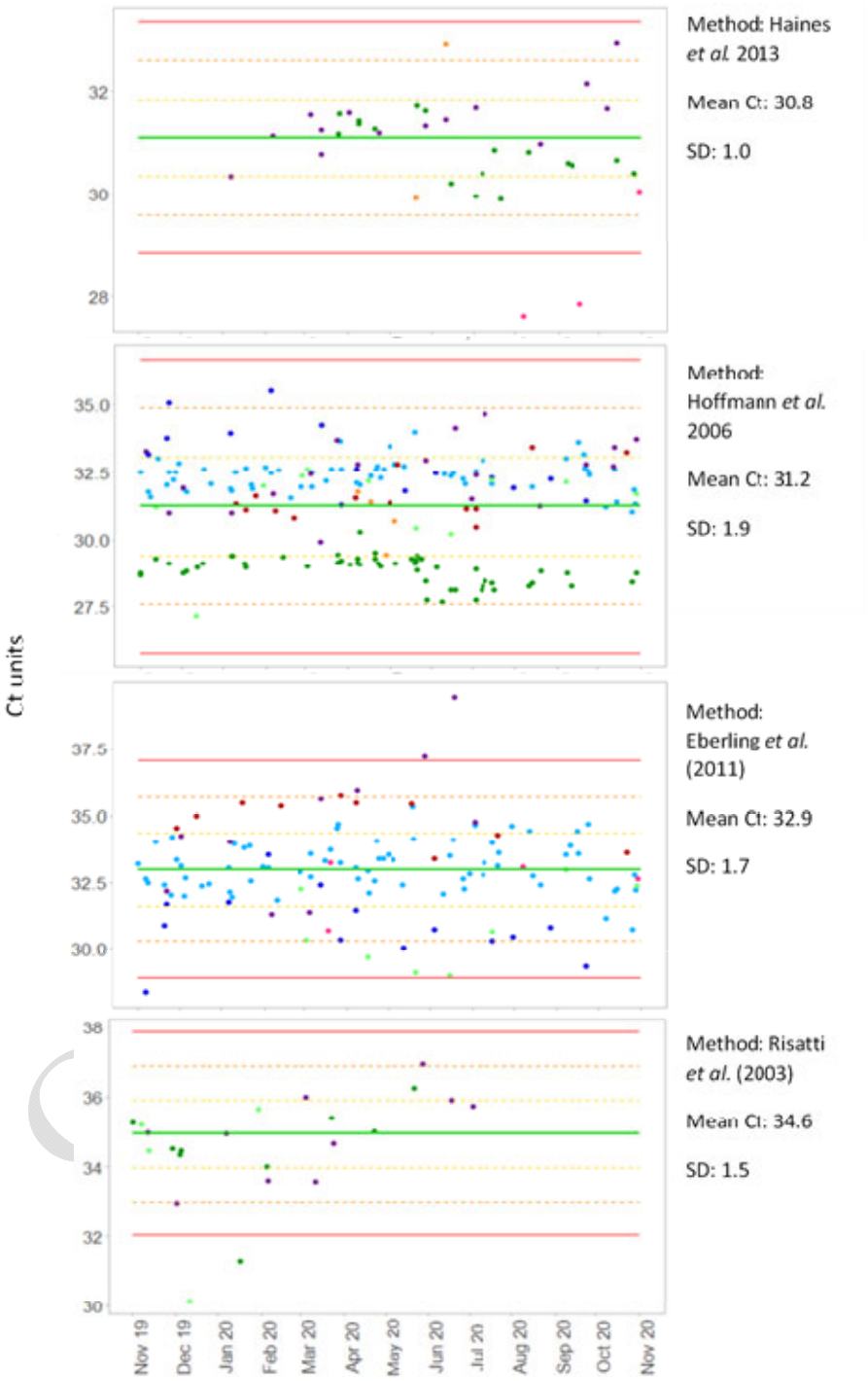


Fig. 1

Results of a network quality control (NQC 17-04) prepared for a network of laboratories for detection of classical swine fever by real-time polymerase chain reaction

The NQC was used as a control every time the assay was performed in each laboratory and data uploaded to a central portal for collation. Accumulated results for the group were analysed (in Excel) for mean and SD, and statistical analysis was then shared with the network at their regular meeting. Data on the NQC 17-04 have been collected since 2017, and accumulated results are shown for the 12-month period November 2019–2020. The results of each assay are displayed as mean Ct value (green solid line) 1SD (yellow dashed line), 2SD (orange dashed line) and 3SD (red solid line). Colour coding is used for each laboratory in the network

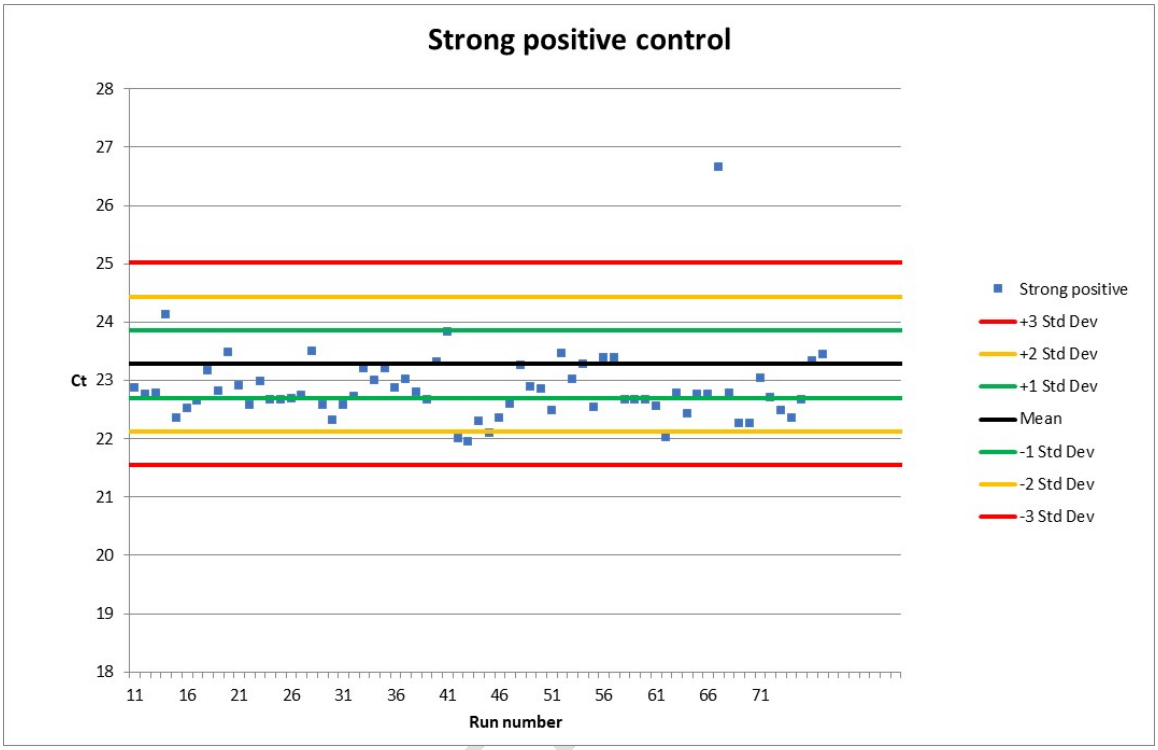
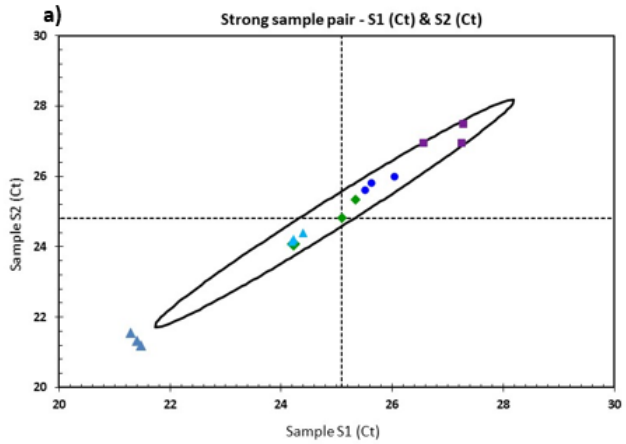


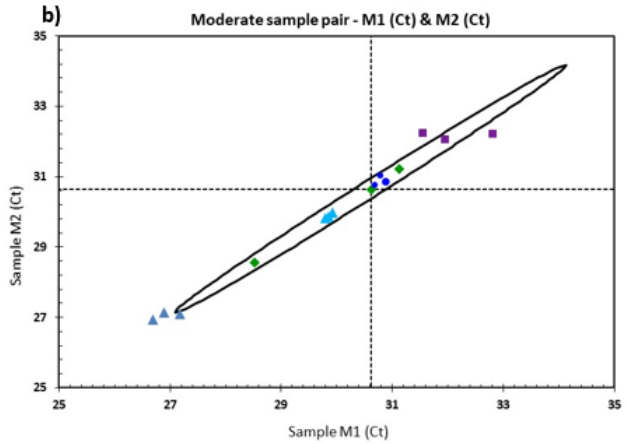
Fig. 2

Example of a Levey-Jennings chart for a strong quantitative real-time polymerase chain reaction extraction control

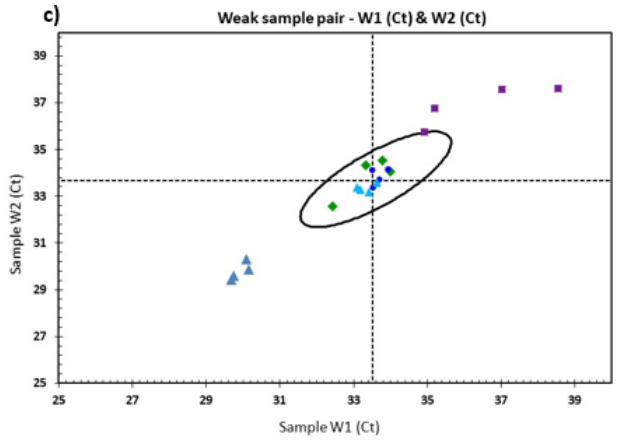
The increase in cycle threshold (Ct) value from run 67 yielded one observation exceeding the mean by +3 SD (rule 1_{s3} Table II). Following this observation, the test run was repeated. The repeated test run shows the control return to within set limits, confirming the random error.



Statistic	Sample S1	Sample S2
No. of Results	15	15
Median	25.09	24.82
Normalised IQR	1.26	1.39
Robust CV	5%	6%
Minimum	21.29	21.19
Maximum	27.29	27.49
Range	6.00	6.30



Statistic	Sample M1	Sample M2
No. of Results	15	15
Median	30.62	30.64
Normalised IQR	1.43	1.49
Robust CV	5%	5%
Minimum	26.69	26.93
Maximum	32.82	32.24
Range	6.13	5.31



Statistic	Sample W1	Sample W2
No. of Results	20	20
Median	33.52	33.66
Normalised IQR	0.78	1.03
Robust CV	2%	3%
Minimum	29.69	29.42
Maximum	38.54	37.59
Range	8.85	8.17

Laboratory 1: dark-blue triangle (▲)

Laboratory 2: diamond (◆)

Laboratory 3: light-blue triangle (▲)

Laboratory 4: circle (●)

Laboratory 5: square (■)

Ct: cycle threshold

CV: coefficient of variation

IQR: interquartile range

Fig. 3

Example of data reported from five different laboratories comparing results for identical samples representing a) strong positives, b) moderate positives, and c) weak positives

The results for each of the sample pairs reported by individual laboratories are plotted on the graph (Youden plot) and a 95% confidence ellipse is placed around the results (based on 2.75 times the standard deviation). The dotted lines represent the median for each sample and the point at which these lines intersect is called the Manhattan median. Points that lie outside the ellipse identify laboratories that produced results which lie outside the 95% confidence limit. Respective summary statistics are captured in tables beside each Youden plot