Monitoring performance of nucleic acid-based diagnostic measurement system users by EQA

H. J. Staines · L. Garcia-Fernandez · R. Pogotheata ·
P. S. Wallace · W. G. MacKay · A. M. Van Loon

Received: 16 October 2008 / Accepted: 3 March 2009 / Published online: 31 March 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract External quality assurance (EQA) programmes allow a laboratory to monitor independently its performance and provide feedback to identify and investigate potential areas of concern. However, both the definition of appropriate ‘performance indicators’ and their statistical analysis vary across EQA providers. Typically, performance scores have an unknown statistical distribution and hence arbitrary thresholds are used to measure performance. In this paper, we introduce a scoring system based on well-known statistical principles. The system is simple, flexible and easy to interpret. It can be used to measure performance for single samples or across a panel so providing useful and meaningful information to participants in EQA programmes. The score is illustrated using the 2005 QCMD Hepatitis B Virus Proficiency programme.

Keywords Quality assessment · Scoring system · Nucleic acid-based diagnostics · Proficiency test data · Statistical methods · Bayesian model

Introduction

External quality assessment (EQA) or proficiency testing (PT) has long been considered the most important way of monitoring laboratory quality. It provides a means of monitoring the entire laboratory practice and procedure, from receipt of sample through to the reporting of patient results. It allows a laboratory to monitor independently performance and also provides a feedback mechanism for identifying and investigating potential areas of concern. As a result, EQA has become an integral part of a laboratory’s quality system requirements, complementing the internal quality control and other quality assurance processes. In addition, with many laboratories now working towards ISO (International Standard Organisation) certification [1] and/or accreditation [2], the need for appropriate EQA has increased substantially. However, although there are specific guidelines [3–5] and general principles which are common to most EQA schemes, there are many different approaches to EQA. Different clinical, analytical, and regulatory goals within different clinical laboratory services and different technologies require variations in the design, implementation, and reporting mechanism of the EQA programme. Hence, it can often be difficult for a clinical laboratory to judge which EQA scheme is best for its specific needs and provide it with the most appropriate way of monitoring its performance. Furthermore, the identification of appropriate criteria or ‘performance indicators’ to monitor and assess a laboratory can be fundamentally difficult. Traditionally, clinical chemistry has led on laboratory quality issues. However, with the advent of new technologies such as nucleic acid-based diagnostics and their rapid transition into the routine clinical laboratory, particularly in microbiology, traditional approaches to quality and EQA are often limited and

H. J. Staines
Sigma Statistical Services Ltd, Balmullo,
KY16 0BJ, St., St. Andrews, Scotland, UK

L. Garcia-Fernandez (✉) · R. Pogotheata
University of Abertay Dundee, Kydd Building,
40 Bell Street, Dundee DD1 1HG, Scotland, UK
e-mail: l.fernandez@abertay.ac.uk

P. S. Wallace · W. G. MacKay · A. M. Van Loon
Quality Control for Molecular Diagnostics, Block 4.1,
Kelvin Campus, West of Scotland Science Park,
Glasgow G20 0SP, Scotland, UK

difficult to apply. Therefore there is a need to adapt current methods or define, develop, and implement new quality rationale for EQAs. One of the greatest challenges is defining suitable performance indicators and monitoring performance in a clinical context as well as being able to differentiate amongst good, average, and poor participating laboratories. Within clinical microbiology, many EQA organisations providing schemes for the molecular diagnosis of infectious diseases have focused on the traditional subjective approach to EQA with peer group review and consensus analysis [6] rather than on defined performance indicators related to specific analytical and clinical parameters. In addition, the lack of high quality reference materials, internationally recognised standards, and suitable control material with which to monitor nucleic acid-based measurement systems (assays) has meant that EQA organisations have been compelled to use consensus-based approaches in order to establish acceptable performance limits.

To allow some internal comparison of laboratory performance, type of technology and method used within an EQA distribution the Quality Control for Molecular Diagnostics (QCMD) organisation has historically used a very simple scoring system for their EQA panels. The EQA panels are designed containing samples with no, low, medium and high microbial concentrations. The panels are distributed to different laboratories worldwide. Laboratories analyse the panels knowing the pathogens but blind to the microbial concentrations. Results obtained by the laboratory are returned to QCMD for analysis. The individual laboratory scores were not reported but used for internal report analysis only. The principal of the EQA programme was non-judgemental and one of self determination. However as regulatory requirements change, increasing numbers of laboratories are asking for a performance score to be provided with each EQA distribution. Although International guidelines for EQA exist, understandably there are no global criteria for what constitutes an acceptable or unacceptable level of performance. Some EQA organisations focus on accuracy and others on repeatability. However, scoring is often on an arbitrary scale and it does not take into consideration the technology, method, and the clinical importance or, implication of a particular result. In addition, within some EQAs the decision on whether to score results for a particular EQA sample is dependant on the correct results obtained by a high proportion of independent pre-testing laboratories, those laboratories which have shown good performance in previous years. In contrast, in other EQAs, an advisory group may directly determine whether the reported results are fit for purpose to use within that EQA distribution. There are also EQAs which use a variation of these approaches.

The QCMD Biostatistics working party evaluated approaches to performance scoring within the QCMD EQA programmes. The primary aim was to establish a suitable mechanism for monitoring laboratory performance that gave an appropriate representation of a laboratory's result, was simple and easy to interpret, had the ability to include cumulative performance scores from different QCMD EQA programmes, and most importantly provided useful and meaningful information to the laboratories which take part in EQA programmes. The approach taken was to utilise a number of defined performance indicators and provide feedback to the participants on various aspects of their performance in relation to these performance indicators within the various QCMD programmes. The principal of the scoring mechanism is introduced here.

Method

We first investigate performance indicators for individual samples within a panel and extend this to an overall performance score for a panel.

We provide performance indicators for the estimation of microbial concentration, sample repeatability and detection of the microbe.

We assume that the EQA panel consists of J samples. The j th, ($j = 1, \dots, J$), sample is assumed to have an estimated target concentration, μ_j , and to be categorised as one of 'Strong positive' (SP), 'Positive' (P), 'Weak positive' (WP) or 'Negative' (N) defined as having high, medium, low and no microbial concentration. We also assume that participant i , ($i = 1, \dots, I$), has reported a microbial concentration of x_{ij} for the sample j on a $\log_{10}$ scale and/or has reported the sample to be 'Negative', 'Not determined' or 'Positive' (nominal (qualitative) measurement).

We start by describing performance indicator for quantitative measurement analysis followed by the description of a performance indicator of within laboratory consistency and microbe detection. Finally, we explain performance indicators for a panel.

Existing quantitative measurement performance indicators for individuals samples

Simple and immediate measures of the performance of participant i for sample j are available based on functions of the error or deviation, $d_{ij} = (x_{ij} - \mu_j)$. Commonly used functions of d_{ij} include the absolute deviation, $|d_{ij}|$, the squared deviation, d_{ij}^2 and the percentage absolute deviation $100|d_{ij}|/\mu_j$ [7]. These metrics can be used as a relative measure to compare laboratories and are easy to compute and interpret. However, their statistical distributions are not

known and so it is difficult to determine limits to identify participants that are performing satisfactorily. An alternative approach is to set acceptance limits, typically $\mu_j \pm 0.5$ [8]. However, for this measure, all values within the range are regarded as equally good and those outside are uniformly bad. For example, if the target value, on a $\log_{10}$ scale, were 3, then all the values between 2.5 and 3.5 are acceptable all others are not. So a value of 2.51 is acceptable and scored the same as the target value of 3 and a value 3.49. The values of 0, 3.51 and 6 are regarded as equally unacceptable, even though their clinical significance may be very different.

Proposed quantitative measurement performance indicator for individual samples

The proposed quantitative measurement performance indicators for individual samples is based on the standardised score of a laboratory's estimated microbial concentration for sample j , x_{ij} , from a set of data with known mean μ_j and known standard deviation σ_j , defined as

$$z_{ij} = \frac{(x_{ij} - \mu_j)}{\sigma_j}.$$

The proposed score for sample j for the i th participant is defined as

$$z_{ij}^* = \min(3, \text{floor}(|z_{ij}|)).$$

The absolute value of z_{ij} is used as it is assumed that an underestimation and overestimation by the same amount indicates equally poor performance. The floor (integer part) function with a maximum of three helps interpretation by participants.

The possible values of z^* are 0, 1, 2 and 3. The score 0–3 are presented to participants as 'highly satisfactory', 'satisfactory', 'unsatisfactory' and 'highly unsatisfactory' and may be visualised with an associated colour code, for example: green, yellow, orange and red respectively.

Note that, in general, the mean, μ_j and the standard deviation, σ_j of the microbial concentration of sample j are not known.

Procedure to calculate quantitative measurement performance indicators for individual samples

Outliers detection

Once participants have submitted their results and data have been cleaned (based on a pre-defined Standard Operation Procedure), participants may be classified into K

mutually exclusive and exhaustive strata (e.g. based on technology used). For those strata with at least five observations the standardised residuals are calculated. Outliers are defined as those values with a standardised residuals with absolute value greater than 3 [9]. When the stratum has less than five observations, outliers are defined as those values greater than 1.5 times the interquartile range from the relevant quartile [10].

Outliers are removed from the data when calculating $\hat{\sigma}_j$, the estimate of the standard deviation, used to obtain the performance indicator z_{ij}^* .

Note that those technology groups with the minimum of 5 datasets are included in this score calculation if and only if the groups contain at least four datasets after detecting and removing outliers.

Mean estimation

The mean μ_j may be estimated by the sample mean, $\bar{x}_j$, known as the consensus value. However, this estimate may be biased towards the mean of the modal measurement system used and may be influenced by poorly performing laboratories [11]. Besides, the mean μ_j may be estimated from the use of a limited number of 'reference' laboratories prior to the distribution of the panel, but this estimate may be inaccurate and biased towards the technology used by them. Hence, the use of a more robust estimate such as a trimmed mean may be more appropriate.

Here, we propose a Bayesian approach to provide a more accurate and appropriate estimate that makes use of a prior estimate target microbial concentration or sample target concentration, ϑ_j , for the j th sample. The prior sample target concentration may be available prior to the panel distribution by the EQA organisation. The proposed estimate is based on the prior target information updated with estimates provided by 'reference' laboratories to obtain the 'posterior information'. This is the distribution around the most likely true concentration target based on the information available.

Bayesian model description The prior information is represented by the distribution of the unknown sample target concentration, μ_j , and the observed information by the estimated sample target concentration by reference laboratories, y_{rj} for the laboratory r with $r = 1, \dots, R$ and sample j . In the proposed performance indicator of individual samples, it is assumed that the prior and observed distributions are normal.

The prior distribution of μ_j is assumed to be $N(\vartheta_j, \tau_j^2)$. The mean ϑ_j is a defined prior target concentration for sample j and the variance τ_j^2 is chosen to be 0.0625 for all samples since this ensures that the 95% of the prior distribution lies within the interval $\vartheta_j \pm 0.5$ [12].

The distribution of y_{rj} is defined as $N(\mu_j, \zeta_j^2)$ where ζ_j^2 is an unknown parameter having an Inverse Gamma distribution with parameters a and b , $IGamma(a, b)$. Since we do not have proper prior information about ζ_j^2 , the parameters a and b are taken to be 0.0001 corresponding to a non-informative prior distribution. Note that other distributions can be used for the unknown ζ_j^2 [13].

The conditional posterior distribution for the target concentration μ_j is the normal distribution [13],

$$N\left(\frac{\zeta_j^2 \vartheta_j + \tau_j^2 \sum_r y_{rj}}{\zeta_j^2 + \tau_j^2 R}, \frac{\tau_j^2 \zeta_j^2}{\zeta_j^2 + R \tau_j^2}\right).$$

Therefore, the proposed estimate for the true value of target concentration is

$$\hat{\mu}_j = \frac{\zeta_j^2 \vartheta_j + \tau_j^2 \sum_r y_{rj}}{\zeta_j^2 + \tau_j^2 R}.$$

Standard deviation estimation

The standard deviation, σ_j , may be estimated by using s_j , the sample standard deviation. Assuming the participants are a random sample of all diagnostics technologies users, then s_j^2 is an unbiased estimate of σ_j^2 .

However, since participants may be classified into K mutually exclusive and exhaustive strata (e.g. based on technology used), a more accurate estimate of the pooled standard deviation, σ_j , may be found by considering the strata sizes and within strata standard deviations as shown below.

If there are n_k participants within stratum k ($k = 1, 2, \dots, K$) and their standard deviation for sample j is s_{jk} , then an unbiased estimate for σ_j^2 is found from

$$\hat{\sigma}_j^2 = \frac{\sum_k (n_k - 1) s_{jk}^2}{\sum_k (n_k - 1)}.$$

This is the value of the mean square for the error in the ANOVA table when the response is the participant's result and the factor levels are the strata.

Depending on the objectives of the assessment, participants' laboratories may be assessed with respect to the estimate target concentration, consensus value or stratum consensus value.

To illustrate how to calculate scores for an individual sample, consider two laboratories, Lab1 and Lab2, with estimated microbiological concentrations of 3.509 and 1.826, respectively. These resulted in standardised residuals of -0.658 and -3.520 , respectively.

Lab2 is detected as an outlier as its standardised residual is outside the interval $(-3, 3)$. The individual score based on the consensus concentration, technology consensus concentration, target sample concentration (estimate

available by the EQA organisation) and the target concentration mean (Bayesian estimate for the sample concentration) are calculated as follows:

- The consensus mean and standard deviation once outliers are removed are calculated as 3.988 and 0.473. Therefore, the score is calculated in the following way:

$$\text{Lab1 } z_{\text{lab1}} = \frac{3.509 - 3.988}{0.473} = -1.013,$$

$$z_{\text{lab1}}^* = 1$$

$$\text{Lab2 } z_{\text{lab2}} = \frac{1.826 - 3.988}{0.473} = -4.571,$$

$$z_{\text{lab2}}^* = 3$$

Alternatively, the target mean of 4 could be used resulting in a z score for Lab1 of -1.038 and a z^* score of 1. The Bayesian estimate for the mean concentration is 4.469 yielding a z score for Lab1 of -2.029 and hence a z^* score of 2. The proposed score is flexible and instead the technology mean and standard deviation can be used. Their estimates are 3.957 and 0.593, respectively. This leads to a z score for Lab1 of -0.755 and thus a z^* score of 0.

Proposed performance indicator of within laboratory consistency

To assess the performance of within laboratory consistency we suggest the following procedure:

- Calculate the difference of a laboratory's results for two samples, d_i .
- Estimate the standard deviation of the differences, $\hat{\sigma}_d$, as we did for calculating the quantitative measurement score.

$$\hat{\sigma}_d^2 = \frac{\sum_k (n_k - 1) s_{dk}^2}{\sum_k (n_k - 1)},$$

where s_{dk} is the standard deviation, of the difference of estimated microbial concentration for the two samples, for the stratum k .

- Calculate the score, $Z_{d_i}^*$, based on the previous score formula, where

$$z_{d_i}^* = \min\left(3, \text{floor}\left[\left|\frac{(d_i - \hat{d})}{\hat{\sigma}_d}\right|\right]\right).$$

where $\hat{d}$ is the known or estimated mean for the difference of microbial concentration for the two samples.

The interpretation of the score is equivalent to the quantitative measurement score.

Repeatability is a special case to monitor laboratory consistency with $d = 0$. We define repeatability as a laboratory producing the same estimate microbial concentration for two identical samples within a panel. Note that repeatability score must not be added to obtain the panel score since it is not independent to the scores of individual samples of the panel. Repeatability score is extra information about the ability to reproduce sample concentration concentrations.

Existing microbe detection performance indicators for individual samples

A common measurement for scoring the performances of participants with nominal (qualitative) response ('Positive', 'Negative') is to assign a score of 2 if the result is correct and 0 if the result is incorrect or not determined. However, the severity of the error is not taken into account (e.g. reporting a negative sample when the sample contains high microbial concentration should be penalised more than reporting a negative sample when the sample contains low microbial concentration).

Proposed microbe detection performance indicator for individual samples

The proposed score for an individual sample ranges from 0 to 3 as defined in Table 1. This performance indicator takes into account the severity of the error.

Proposed performance indicator for a panel

The proposed performance indicator for an individual sample ranges from 0 (highly satisfactory) to 3 (highly unsatisfactory). One measure of overall performance for a panel is to sum a participant's score for those samples where a value is reported. The distribution of this score is not known and will vary according to the number of samples reported. Participants are classified using the method outline below.

Table 1 Penalty score table for microbe detection analysis

Sample category	Participant's result		
	Negative	Not determined	Positive
Strong positive	3	3	0
Positive	2	2	0
Weak positive	1	1	0
Negative	0	3	3

Quantitative measurement results

Assuming normality and independence, the proposed score for an individual sample takes the value 0–3 with probabilities 0.683, 0.272, 0.043 and 0.002, respectively. J columns of ten thousand Monte Carlo simulations from the above probability mass function are found to replicate 10000 virtual participants. The frequencies for these sums are found. For consistency with the scoring for individual samples, participants that reported J samples are given score 0 (classified as 'highly satisfactory') if their sum is in the smallest 68.3% of the simulated values, score 1 ('satisfactory') in the next 27.2%, score 2 ('unsatisfactory') in the following 4.3% and score 3 ('highly unsatisfactory') in the highest 2%. Table 2 gives the range of total scores corresponding for each panel score for given number of samples. For example, a participant with a total score of six or seven from seven samples would be scored 2 (unsatisfactory).

Microbe detection results

A similar approach is taken for the panel score for microbe detection data. For each sample, j , we observe a proportion p_j of participants with the correct response. Ten thousand observations from a Bernoulli trial with probability $(1 - p_j)$ are simulated. These are then multiplied by 3 (for a strong positive or negative sample) or 2 (for a positive sample). The results for each sample are summed and the panel score ranging from 0 (highly satisfactory) to 3 (highly unsatisfactory) is based on the cumulative 68, 95,

Table 2 Panel score table up to 15 samples per panel

Number of samples	Panel score			
	0	1	2	3
1	0	1	2	3
2	0	1	2–3	4+
3	0–1	2	3–4	5+
4	0–1	2–3	4–5	6+
5	0–1	2–4	5–6	7+
6	0–2	3–4	5–7	8+
7	0–2	3–5	6–7	8+
8	0–3	4–5	6–8	9+
9	0–3	4–6	7–8	9+
10	0–4	5–6	7–9	10+
11	0–4	5–7	8–10	11+
12	0–4	5–7	8–10	11+
13	0–5	6–8	9–11	12+
14	0–5	6–8	9–12	13+
15	0–6	7–9	10–13	14+

99 and 100% cut off points for the sum of the scores for all samples in the panel. For example, a participant that was within the 68% group of participants with the lowest summed score would have a panel score of 0.

Results and application

Here, we present an application of the proposed performance indicators to the 2005 QCMD Hepatitis B Virus Proficiency Programme [14]. The Panel composition is shown in Table 3. Prior to the distribution of the panel, a total of 7 panels were analysed by independent ‘reference’ laboratories. A total of 122 data sets were received (all had microbe detection data whilst 101 also included microbiological concentration estimates) from 116 participants within 27 countries.

The datasets were classified per technology groups used to analyse the panel: conventional commercial PCR (CC, $n = 38$), Conventional In-house PCR (CIH, $n = 12$), Real time Commercial PCR (RTC, $n = 39$), Real time In-house PCR (RTIH, $n = 24$), bDNA (bDNA, $n = 7$) and Hybrid Capture (HC, $n = 2$). However, the total number of datasets reporting quantitative measurement data per technology group varied depending on the sample analysed.

Quantitative measurement analysis

We present a summary of the quantitative measurement score obtained with respect to consensus and technology group consensus sample concentration.

Note that the negative samples and values reported as outside the detection limits of the measurement systems were excluded from the quantitative measurement analysis. Additionally, all laboratories except those with results outside the limit of detection of the measurement system

are scored even if they were not included for the calculation of the standard deviation.

Score with respect to consensus sample concentration

The consensus mean $\bar{x}_j$ and standard deviation, $\hat{\sigma}_j$, for each positive sample were calculated (datasets provided by laboratories using HC technologies and outliers were not included for the calculations). We used four technology groups: CC, RTC, RTIH and bDNA, to estimate the standard deviation for samples HBV01, HBV02, HBV04, HBV06 and HBV08. However, we did not have enough datasets (≥ 5) to take into account the bDNA group for samples HBV03 and HBV05 since some estimates were outside the limit of detection for the measurement system.

Table 4 shows the estimated mean and standard deviation of the log microbial concentration and the frequency of z^* scores for each sample with respect to the consensus.

Score with respect to technology consensus mean

We calculate the consensus mean for each technology group with at least four observations once outliers have been removed. Each laboratory, associated with a technology group, is scored with respect to the consensus mean and the estimated standard deviation of that technology group. CIH and HC technology groups do not satisfy the requirement of at least five datasets before and four datasets after outliers removal. Therefore, laboratories using these technologies are excluded for scoring with respect to their technology group.

Table 5 shows the estimated mean and standard deviation of the log microbial concentration and the frequency of z^* scores for each sample by technology.

Scoring performance within laboratory consistency

As an illustrative example, we consider sample HBV01 and HBV04, both of which are subtype D with targets 5.00 and 4.00 $\log_{10}$ copies/mL, respectively. The consensus mean and standard deviation of the difference after removal of outliers are 0.942 and 0.230, respectively. Therefore, the difference, z value and the score for laboratory consistency of a laboratory with observations 4.954 and 3.923 for samples HBV01 and HBV04, respectively are:

$$d = 1.031, \quad z_d = \frac{1.031 - 0.942}{0.230} = 0.387 \text{ and } z_d^* = 0.$$

Microbe detection analysis

Table 6 shows the total number of datasets per technology group and percentage of datasets with a score of 0, 1, 2 and 3 per sample.

Table 3 2005 QCMD HBV Panel composition

Sample	Sample subtype	QCMD concentration unit	Defined sample status
HBV06	D	6.00	SP
HBV08	A	5.00	P
HBV01	D	5.00	P
HBV02	A	4.00	P
HBV04	D	4.00	P
HBV05	A	3.00	WP
HBV03	D	3.00	WP
HBV07	N/A	0.00	N

SP Strong positive, P positive, WP weak positive, N negative

QCMD concentration unit = $\log_{10}(N)$ where N is the number of copies in 1 mL sample solution

Table 4 Summary Laboratories' Score with respect to estimated consensus concentration

	Consensus		Conventional commercial PCR (CC) <i>n</i> = 32					Conventional in-house PCR (CIH) <i>n</i> = 4				
	Mean	SD	0	1	2	3	LOD/NR ^a	0	1	2	3	LOD/NR ^a
HBV06	5.702	0.607	16	3	1	0	12	3	0	1	0	0
HBV08	4.887	0.511	29	2	0	0	1	0	1	2	1	0
HBV01	4.789	0.576	31	0	1	0	0	1	1	1	1	0
HBV02	3.988	0.473	28	4	0	0	0	2	0	2	0	0
HBV04	3.834	0.544	29	2	1	0	0	0	1	2	1	0
HBV05	2.952	0.576	29	2	0	0	1	1	1	0	0	2
HBV03	2.879	0.453	28	3	0	0	1	1	0	3	0	0
	Consensus		Real time commercial (RTC) <i>n</i> = 35					Real time in-house (RTIH) <i>n</i> = 20				
	Mean	SD	0	1	2	3	LOD/NR ^a	0	1	2	3	LOD/NR ^a
HBV06	5.702	0.607	22	9	3	1	0	14	4	0	2	0
HBV08	4.887	0.511	21	10	2	2	0	13	2	3	2	0
HBV01	4.789	0.576	21	9	1	4	0	14	2	2	2	0
HBV02	3.988	0.473	21	10	2	2	0	12	4	2	2	0
HBV04	3.834	0.544	20	11	0	3	1	12	6	1	1	0
HBV05	2.952	0.576	19	8	6	0	2	12	2	1	0	5
HBV03	2.879	0.453	16	12	2	1	4	9	5	1	1	4
	Consensus		bDNA <i>n</i> = 7					Hybrid capture (HC) <i>n</i> = 2				
	Mean	SD	0	1	2	3	LOD/NR ^a	0	1	2	3	LOD/NR ^a
HBV06	5.702	0.607	7	0	0	0	0	2	0	0	0	0
HBV08	4.887	0.511	7	0	0	0	0	0	0	0	0	2
HBV01	4.789	0.576	7	0	0	0	0	0	0	0	0	2
HBV02	3.988	0.473	7	0	0	0	0	0	0	0	0	2
HBV04	3.834	0.544	7	0	0	0	0	0	0	0	0	2
HBV05	2.952	0.576	0	1	0	0	6	0	0	0	0	2
HBV03	2.879	0.453	0	0	0	0	7	0	0	0	1	1

CC commercial PCR, CIH conventional in-house PCR, RTC real time commercial PCR, RTIH real time in-house PCR, HC hybrid capture

^a LOD/NR: result reported as lower limit detection or upper limit detection/no value or not result reported

Panel score

QCMD only provides a panel score to those laboratories that report the estimated concentrations for all positive samples. Thus, panel scores for quantitative measurement analysis were found for the 56 laboratories which returned estimates for all seven positive samples. Of these, 47 datasets received a panel score 0, 2 datasets obtained score 1, 3 datasets obtained score 2 and 4 datasets obtained score 3.

Discussion

EQA is an important mechanism for laboratories to measure their performance of the entire process. However,

performance indicators used by EQA providers varies with limits set for good performance arbitrary often without documented statistical justification. Typical performance indicators consider the detection of a microbe, repeatability and the accuracy of microbial concentration estimates. In this section, we consider each of these in turn concentrating on their scoring schemes.

Microbe detection

In contrast to many EQA schemes, in this paper, we propose a score for detecting a microbe within the sample that takes into account the microbial concentration. This penalises the failure to detect a sample with a high microbial concentration more than failing to detect a low microbial concentration. Clearly, defining the categories

Table 5 Summary laboratories' score with respect to estimated technology consensus concentration

	Technology		CC <i>n</i> = 32					Technology		bDNA <i>n</i> = 7				
	Mean	SD	0	1	2	3	LOD/NR ^a	Mean	σ_j	0	1	2	3	LOD/NR ^a
HBV06	5.721	0.496	6	12	2	0	12	5.940	0.035	5	2	0	0	0
HBV08	5.068	0.241	27	3	1	0	1	5.161	0.092	6	1	0	0	0
HBV01	4.967	0.281	28	1	2	1	0	5.004	0.069	5	2	0	0	0
HBV02	4.121	0.253	26	3	3	0	0	4.163	0.140	4	3	0	0	0
HBV04	4.017	0.291	26	5	1	0	0	3.988	0.125	5	2	0	0	0
HBV05	3.102	0.257	29	2	0	0	1	–	–					
HBV03	2.976	0.255	14	16	1	0	1	–	–					

	Technology		RTC <i>n</i> = 35					Technology		RTIH <i>n</i> = 20				
	Mean	SD	0	1	2	3	LOD/NR ^a	Mean	σ_j	0	1	2	3	LOD/NR ^a
HBV06	5.637	0.720	24	9	1	1	0	5.676	0.574	13	4	1	2	0
HBV08	4.874	0.679	22	11	1	1	0	4.700	0.568	14	4	0	2	0
HBV01	4.697	0.799	24	7	3	1	0	4.641	0.563	13	4	1	2	0
HBV02	3.957	0.593	24	9	0	2	0	3.788	0.577	14	3	2	1	0
HBV04	3.763	0.716	23	8	2	1	1	3.806	0.609	12	7	0	1	0
HBV05	2.794	0.754	22	11	0	0	2	2.937	0.601	12	2	1	0	5
HBV03	2.836	0.564	19	11	0	1	4	2.855	0.525	9	6	0	1	4

CC commercial PCR, RTC real time commercial PCR, RTIH real time in-house PCR

^a LOD/NR: result reported as lower limit detection or upper limit detection/no value or not result reported

Table 6 Summary microbe detection laboratories' score

	Defined Status	CC, <i>n</i> = 38				CIH, <i>n</i> = 12				RTC, <i>n</i> = 39			
		0	1	2	3	0	1	2	3	0	1	2	3
HBV06	SP	38	0	0	0	12	0	0	0	39	0	0	0
HBV08	P	38	0	0	0	12	0	0	0	39	0	0	0
HBV01	P	38	0	0	0	12	0	0	0	39	0	0	0
HBV02	P	38	0	0	0	12	0	0	0	39	0	0	0
HBV04	P	38	0	0	0	12	0	0	0	38	0	1	0
HBV05	WP	36	2	0	0	8	4	0	0	37	2	0	0
HBV03	WP	35	3	0	0	10	2	0	0	34	5	0	0
HBV07	N	38	0	0	0	12	0	0	0	38	0	0	1

	Defined Status	RTIH, <i>n</i> = 24				bDNA, <i>n</i> = 7				HC, <i>n</i> = 2			
		0	1	2	3	0	1	2	3	0	1	2	3
HBV06	SP	24	0	0	0	7	0	0	0	2	0	0	0
HBV08	P	24	0	0	0	7	0	0	0	0	0	2	0
HBV01	P	23	0	1	0	7	0	0	0	0	0	2	0
HBV02	P	23	0	1	0	7	0	0	0	0	0	2	0
HBV04	P	23	0	1	0	7	0	0	0	0	0	2	0
HBV05	WP	18	6	0	0	3	4	0	0	0	2	0	0
HBV03	WP	19	5	0	0	2	5	0	0	1	1	0	0
HBV07	N	22	0	0	2	6	0	0	1	1	0	0	1

CC commercial PCR, CIH conventional in-house PCR, RTC real time commercial PCR, RTIH real time in-house PCR, HC hybrid capture

SP strong positive, P positive, WP weak positive, N negative

strong positive, positive and weak positive is open to debate. This may be done using a combination of clinical relevance and by considering the proportion of participants detecting the microbe (e.g. at least 95% of participants must detect the microbe for the sample to be defined as a strong positive).

The individual samples scores for the proposed measure are summed to form a panel score. We suggest that Monte Carlo methods are used to simulate 10000 virtual users using the probabilities of detecting a microbe estimated from participants. These are summed and used to distinguish between good, average and poorly performing laboratories.

Estimation of microbial concentration

In contrast to many EQAs that exclusively use a consensus mean, we suggest a Bayesian approach to estimate the target value. The Bayesian estimate is the value suggested by the EQA provider (from internal investigations and previous panels) amended by estimates from reference laboratories. Although these too may contain measurement system bias, care can be taken to ensure a range of measurement systems are covered by the reference laboratories.

We suggest a measure that uses a more accurate measure of the target value (see above) and a function of the laboratory's absolute deviation from the target. We find the score by dividing the absolute deviation by the standard deviation once the systematic error caused by different technologies has been removed. To ease interpretation by laboratories the integer value is found and capped at 3. We have shown (data not reported) that once outliers are removed, the $\log_{10}$ values from participants approximately follow a normal distribution. Hence, we can apply well-known statistical results to this simple, robust measure. It should be noted that good laboratories score low marks for the proposed score but high marks for many existing measures.

The scoring system described here allows for evaluation by technology type and therefore reduces the impact of technology bias.

The proposed measure is the first to use Bayesian techniques to estimate the target value and to incorporate the variation attributable to different examining system technologies. It is flexible as the target value may be replaced by the technology consensus. For this application the laboratories are being assessed on how close their estimate is to the mean of other users of the same technology rather than the target. Hence, using this measure, potential measurement system bias is ignored.

Repeatability

Some EQA providers concentrate on repeatability. For example a laboratory may be regarded as adequate if its

reported difference from two samples is within 0.5 of the median difference of all participants. Note that a laboratory that under (or over) estimates both samples by the same amount would be regarded as adequate in this example. This would be the case even if the estimates were far from the true value. Note also that no use is made of the different technologies used.

Our flexible proposed scoring system could be easily adapted to assess repeatability by using the difference between two measurement systems. For identical samples the target difference is zero. For non-identical samples the target difference can be estimated from the difference between the Bayesian estimates of the two samples.

The scores from individual samples of either the estimated microbial concentration or repeatability measures can be summed across the positive samples in the panel. The use of Monte Carlo simulation using the well-known properties of the normal distribution can be used to classify laboratories as highly satisfactory to highly unsatisfactory. This feedback is more informative and statistically rigorous than currently available.

Drawbacks of the proposed approach

The proposed score provides a flexible, statistically rigorous metric to assess laboratory performance for examining system nucleic acid-based diagnostic users. There are some drawbacks when considering quantitative measurement results.

First, the score requires participants to provide their estimate of the microbial concentration. Sometimes, participants report values as outside the detectable limits of the measurement system they used. These have been ignored for the purposes of this paper. Possible approaches to this problem include using censored value techniques or to replace the value by either the limit of detection or half this value.

Second, the score is only appropriate for positive samples. The score assumes normality, which is almost certainly not valid for reported false positives from negative samples. The low frequency of these makes testing for normality suspect.

Conclusions

The application of these performance indicators to the 2005 QCMD Hepatitis B Virus Proficiency programme highlights the flexible use and desirable properties of the proposed scoring system for assessing various aspects of laboratory performance.

The proposed scoring system for assessing laboratory performance cannot only be applied to the field of

microbiology but it can be applied to other EQA programmes from the field of laboratory medicine to chemistry where the normality assumption of the data is verified.

The Monte Carlo approach for panel scores can be used in any EQA programme where numeric performance scores are given to individual samples.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. ISO 9001:2000 Quality management systems: requirements
2. ISO/IEC 15189:2007 Medical laboratories
3. ISO/IEC Guides 43-1 (1997) Proficiency testing by inter laboratory comparisons. Part1: development and operations of proficiency testing scheme
4. ISO/IEC Guides 43-2 (1997) Proficiency testing by inter laboratory comparisons. Part2: selection and use of proficiency testing schemes by laboratory accreditation bodies
5. ILAC-G13 (2000) Guidelines for the requirements for the competence of providers of proficiency testing schemes
6. CEN TC 140 prEN 14136. Use of EQA schemes in the assessment of the performance of in vitro diagnostic examination procedures
7. Westgard JO (2004) Clinical quality vs analytical performance: what are the right targets and target values? *Accred Qual Assur* 10:10–14
8. Wang LN et al (2008) External quality assessment on detection of hepatitis c virus rna in clinical laboratories of china. *Chin Med J* 121(11):1032–1036
9. Zar JZ (1999) Biostatistical analysis, 4th edn. Prentice Hall, Englewood Cliffs
10. Upton G, Cook I (2002) A dictionary of statistics. Oxford University Press, USA
11. Wong SK (2005) Evaluation of the use of consensus values in proficiency testing programmes. *Accred Qual Assur* 10:409–414
12. Valentine-Thon E, van Loon AM, Schirm J, Reid J, Klapper PE, Cleator GM (2001) European proficiency testing program for molecular detection and quantification of Hepatitis B virus DNA. *J Clin Microbiol* 39:4407–4412
13. Gelman A, Carlin JC, Stern H, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman & Hall, New York
14. Quality Control for Molecular Diagnostics (QCMD). <http://www.qcmd.org>