Special issue: External Quality Assessment in Laboratory Medicine

Review

Demystifying EQA statistics and reports

Wim Coucke*, Mohamed Rida Soumali

Scientific Institute of Public Health section Quality of Medical Laboratories, Brussels, Belgium

*Corresponding author: wim.coucke@wiv-isp.be

Abstract

Reports act as an important feedback tool in External Quality Assessment (EQA). Their main role is to score laboratories for their performance in an EQA round. The most common scores that apply to quantitative data are Q- and Z-scores. To calculate these scores, EQA providers need to have an assigned value and standard deviation for the sample. Both assigned values and standard deviations can be derived chemically or statistically. When derived statistically, different anomalies against the normal distribution of the data have to be handled. Various procedures for evaluating laboratories are able to handle these anomalies. Formal tests and graphical representation techniques are discussed and suggestions are given to help choosing between the different evaluations techniques. In order to obtain reliable estimates for calculating performance scores, a satisfactory number of data is needed. There is no general agreement about the minimal number that is needed. A solution for very small numbers is proposed by changing the limits of evaluation.

Apart from analyte- and sample-specific laboratory evaluation, supplementary information can be obtained by combining results for different analytes and samples. Various techniques are overviewed. It is shown that combining results leads to supplementary information, not only for quantitative, but also for qualitative and semi-quantitative analytes.

Key words: external quality assessment; statistics; Q-score; Z-score

Received: February 15, 2016 *Accepted: July 29, 2016*

Introduction

Reports created by External Quality Assessment (EQA) providers serve as a major feedback tool towards the participating laboratories. They support the pedagogic role of EQA and are often used by auditors to follow up laboratory quality, certainly in the light of eventual accreditation (1–4). Different EQA providers summarize the statistical evaluation and their findings in various types of reports.

In a first instance, participating laboratories should receive, as soon as possible after an EQA round closing, a confidential individual report detailing their own performances. The report should be as clear and comprehensive as possible and contain the assigned values for each of the parameters that were included, limits of acceptability and evaluation for each of the laboratory's result. Ideally, it would contain additional information to support evaluation, like the number of laboratories involved in the evaluation and details about the distribution of data reported by all the participants. As such, the report allows the participating laboratory to compare its results for each analyte with those of other participants (1,5–9). In addition to individual reports for each participant, summary reports containing general and anonymized information on method performance, variability and bias for different analytes could be included at the end of each round. Periodic reports can be made as well to highlight the most striking evidence that is found for different EQA rounds together (7). This manuscript focuses on the feedback reports of individual laboratories and gives an overview of var-

http://doi.org/10.11613/BM.2017.006

ious relevant statistical evaluation techniques of reported data, without aiming at describing the entire range of performance assessment systems.

Because of large differences in EQA scheme design, evaluation procedures vary widely and depend on, among others, choices made for determining the assigned value, commutability of control samples or the way in which laboratories report their results in routine. Commonly, EQA in the clinical field asks laboratories to analyse the samples as if they were routine samples and hence, produce mostly one value for a certain analyte without reporting measurement uncertainty (10). For many analytes determined in the clinical laboratory, reference method-based assigned value setting is not possible. Due to a complex matrix like whole blood or serum, which is pooled for large-scale distribution and subject to procedures to enhance sample stability, samples are altered. Consequently, samples are often not commutable, i.e. the differences between methods that they demonstrate do not reflect the differences that are observed for routine samples (10). Commutable samples enable EQA providers to derive more information from an EQA round than non-commutable samples, like harmonization between methods (4,11). If commutability cannot be assessed, the only way to evaluate laboratories is with respect to their own peer groups. Peer groups consist of laboratories whose measurement procedures are equal or so similar that they are expected to have the same result and matrix-related bias compared to other methods. Peer group evaluation provides valuable information to assess quality, verifying that a laboratory is using a measurement procedure in accordance to the manufacturer's specifications and to other laboratories using the same technology, but cannot assess laboratory or method accuracy (4,11). Commutable samples on the other hand, give insights into the bias and accuracy that reflect analytical performance for routine samples.

In order to help interpreting an EQA result that is out of consensus, EQA providers are encouraged to write advice for poor performers in the report (8). Laboratories should always follow up any unacceptable EQA result by a root cause analysis and document corrective actions (12). In addition, when interpreting EQA results, laboratories should not forget that results within the acceptance range may still be linked to a problem in the laboratory, for example when they are close to the acceptance limits or when successive Z- or Q-scores are all positive or negative (11).

Building performance statistics

Laboratories are marked for an out of consensus result if they report a value that is too far from the assigned value and hence prior to any interpretation, the EQA provider must determine the assigned value and a range of acceptable values around it (1,8,11,13). Criteria for defining the ranges for acceptability are extremely important. Ranges that are too wide will not allow detecting laboratories with poor performance, while a satisfactory performance will be wrongly flagged if the ranges are too strict (7). It is also very important that acceptability criteria are reliable, or laboratories may lose confidence in the scheme.

The comparison with acceptability ranges is often condensed in two different scores: Z-scores and Q-scores.

A simple evaluation technique consists of calculating Q-scores. They consist of the relative difference between the value reported by the laboratory and the assigned value:

$$\text{Q-score} = \frac{\text{reported value} - \text{assigned value}}{\text{assigned value}}$$

The Q-score is often presented as a percentage and compared with a maximal allowable deviation (6,8,13,14). The limit of acceptability is often considered as the 'fitness for purpose', meaning that a result within the limits of acceptability is 'fit for purpose', or better: 'fit for intended use'. It is important to specify such purpose, which should be derived from external requirements (5,15). External quality assessment providers for clinical laboratories usually adopt the approach of analytical performance specifications (16). The approach includes requirements derived from specific studies or general studies like biological variability, and in a second instance, state of the art performance criteria as well.

Another type of score is the Z-score. It is the difference between the value reported by the laboratory and the assigned value, corrected for the variability:

$$\text{Z-score} = \frac{\text{reported value} - \text{assigned value}}{\text{standard deviation}}$$

If the distribution of the data reported by well performing laboratories approaches a normal distribution, Z-scores follow a standard normal distribution and the percentage of Z-scores that are beyond extreme values can be calculated exactly: 4.6% and 0.27% of the Z-scores will have an absolute value greater than 2 and 3, respectively. Hence, a very small minority of well performing laboratories have Z-scores larger than 2 and even fewer have Z-scores greater than 3. That is why often a Z-score with absolute value lower than 2 is considered as acceptable, between 2 and 3 as questionable and unsatisfactory when it is larger than 3 (3). Because Z-scores are standardized scores, they can be compared between all analytes (8).

As can be seen from the formulas to calculate Q- and Z-scores, they both include an estimate of the assigned value and Z-scores also need an estimate of the variability of the data, expressed as a standard deviation.

Calculating performance scores for quantitative tests: one sample, one parameter

The evaluation of a laboratory in an EQA round is basically an assessment of how well an analyte has been measured in a certain sample. Before calculating any score, EQA providers should examine the reported data and screen them for anomalies that jeopardize a correct evaluation. Ideally, the reported data would be normally distributed. In practice however, EQA providers cannot ensure this assumption and have to check the data for anomalies, of which different types may occur. The most common are bimodality, skewness and outliers.

Bimodality occurs when the data consists of a collection of small groups with different central values. Skewness occurs when the data are not centrally located around their mean, i.e. there is an increased proportion of extremely large or small data. Outliers are probably the most common anomaly. Mostly, outliers are data that are far from the bulk of the data, i.e. the process that produced them is not like the process that produced other data. The process may be out of range, like, for example, showing a systematic deviation or an increased variability, or the outlier could be caused by an extra-analytical mistake, like a clerical error or sample identification mistake. Skewness can be detected by means of graphical exploration of the data and data transformation; like a log- or square root transformation. In most cases, it helps to make the data more symmetrical. In case of bimodality, several statistical tools are available to detect the different subgroup. They rely on kernel density estimation, which is a nonparametric technique to estimate the probability density function from the data and serves excellently for identifying modes. Some use solely kernel density estimation for identifying modes, others extend this technique by a method called bootstrapping (17). It is a method that is based on resampling and aims at estimating the behaviour of the distribution's parameters in order to find the largest mode (18–20). The statistical procedures for handling bimodality and skewness should be applied by the EQA organizer between the deadline for reporting results and the creation of feedback reports. Once the EQA provider has validated these procedures, they remain preferably unchanged over time.

In the following sections, it is assumed that bimodality and skewness have been dealt with either by using homogeneous, unimodal data, or by transformation and that the statistical techniques only have to deal with outliers.

Outlier removal

Unfortunately, the rule that identifies outliers with 100% certainty does not exist. Even more, the detection of outliers has various flaws, like masking and swamping. Masking means that an outlier is not detected by the presence of another outlier, swamping means that a non-outlying observation is falsely indicated as an outlier (3,21,22).

Three tests are commonly used for outlier detection of EQA data: the Hampel outlier test, Grubbs test and Dixon test. The Hampel and Grubbs tests compare the difference between an extreme value and the centre of the data with the variability of the data and identify the extreme value as an outlier if the ratio is too large. The Dixon test looks at the difference between the two most extreme values and an estimator of scale to identify outliers. The three tests can work with a specified alpha, i.e. the probability that value is wrongly marked as outliers, which should be kept as low as possible, like 0.05. For relative small data series (N < 15), a higher value of alpha could be adopted. Recently, the Hampel and Grubbs tests have been proposed as preferable in comparison to the Dixon test (23–25) with the Grubbs test able to handle also small data series, from six data points on (25).

It should be noted that indicating outliers and marking them as "out of consensus results" does not go as far as calculating performance scores, like Z- or Q-scores. Q- and Z- scores can be calculated by identification and removal of outliers prior to calculation of assigned/target value and descriptive statistics, followed by calculation of individual Q and Z scores for all participants, whether outliers or not. Outlier participants should still receive scores even though their results are excluded from calculation of the target value.

Determining the assigned value

Several ways exist to set or determine the assigned value. A first group of assigned value setting possibilities are rather chemical: adding amounts of pure analyte to a sample matrix containing none, certified reference materials with assigned values determined by formulation or analysis with definitive methods or reference values determined by analysis that are traceable to reference standards. In this case, commutability should be assured as well (2,6,8,11,13,14). Other methods rely on statistics: consensus values from reference laboratories that use the best available methods, or from participants (6,8,13,14). It has been reported that over 90% of the programmes rely on consensus values (2). There are numerous methods to assess the assigned value based on reported results and all of them attempt to accommodate for the most common anomaly that may endanger a correct estimation of the assigned value: outliers.

The influence of outliers on the estimation of the central value may be significant even when groups are unimodal and symmetrical. When the classical average is used, outlier detection tests, as described in the previous section, should be applied to identify and exclude outliers before the average. Another possibility is to use techniques that attempt to find a correct estimate of the assigned value in presence of outliers. Estimators obtained by these techniques are called robust estimators, since they are not, or almost not, influenced by outliers. Two criteria play a role in the evaluation of these robust estimators: breakdown point and efficiency. The breakdown point can be seen as the proportion of the data that could be infinite without influencing the estimate to be infinite. Hence, the higher the breakdown point, the more outliers may be present in the data before a clear effect on the estimated assigned value is visible. Efficiency reflects the uncertainty of the estimator: high-efficient estimators are very certain. In general, high breakdown point and high efficiency are antagonistic criteria, i.e. high breakdown point is associated with low efficiency. For example, the classical average has a high efficiency, but a very low breakdown point. The kernel density-based estimation of the mode on the other hand, has a very high breakdown point, but low efficiency.

One of the most widely used estimators of the assigned value is the median (7). It is simply the middle value when the reported values are sorted from smallest to largest. Medians have a very high breakdown point, but exhibit a low efficiency. Other estimators exist that have an acceptable breakdown point and have a better efficiency than the median, like the estimator from Algorithm A from the ISO 13528 (13). Originally described by Huber as the H1.5 algorithm (26), this algorithm starts with an estimation of the central location, and subsequently reduces the influence of outlying results by winsorization, i.e. changing values outside an interval by the outer values of the interval (27).

In addition to the well-established estimators, some less known estimators merit mentioning as well. In fact, there is a family of central location estimators that offer solutions for the following algorithm:

The parameter θ is the estimator of location for which $\Sigma_{i=1}^{n} |x_i - \theta|^p$ is minimal, where by x_i are the n data points and p is a predefined value (28). For a certain value of p, there is only one value of θ that minimizes this sum for a given data series. This value is called the least power (Lp) estimate. It is interesting to know that the classical average is obtained by setting p to 2, and the median is obtained by setting p to 1. Because classical average is strongly biased towards outliers but has a very high efficiency, while the median has a low efficiency, it may be interesting to think of an intermediate estimator. This estimator is found by setting p to 1.5, and is called the L1.5-estimator. It is more efficient than the median and is less influenced by outliers than the average.

Another estimator is the MM-estimator, which should have a very low bias towards outliers and is more efficient than the other estimators that are presented here (29,30). Its calculation is relatively complicated though.

Determining the standard deviation

Similar to the case of the assigned value, different ways exist to determine the standard deviation and the EQA provider adopts its own procedure for its determination (6). They belong to two distinct classes. The first class contains the parameters that are fixed beforehand. They may be a value derived from a perception of how laboratories should perform, legislative documents, a small-scale trial from a model of precision, like the Horwitz curve (1,7,8,13,31). The latter however is rarely applied in EQA schemes for clinical laboratories. If historic data are available, the standard deviation could be derived from the assigned value, for example by means of the characteristic function (32,33), which is a mathematical relation to estimate the standard deviation based on the assigned value:

$$\text{SD} = \sqrt{\alpha^2 + \beta^2 \times (\text{assigned value})^2}$$

where α and β are to be estimated from the historical data by means of non-linear regression. The coefficients α and β have a different meaning in explaining the standard deviation. The parameter α principally explains the standard deviation at low concentrations, while the parameter β affects the standard deviation at higher concentrations and approaches the coefficient of variation (CV) when β is low or the concentration is high.

The second class contains the estimates of standard deviation that are based on the reported results.

Since reported EQA data may have outliers, the classical estimate of standard deviation should only be used after elimination of outliers, as identified by the Dixon or preferentially the Huber or Grubb test, since the presence of only a few outliers inflate it and make it unreliable.

EQA providers could also rely on robust estimators for the standard deviation. The ISO 13528 standard proposes Huber's M-estimator H1.5 (called algorithm A), also for the estimate of variability (13). Other methods propose the robust Qn estimator, which is expected to be more efficient, but loses reliability in case the same value occurs more than once in the data set (34,35).

Another estimator that is easy to calculate is based on the interquartile range (IQR), in which the standard deviation is estimated by dividing the IQR by 1.349 (7,36,37).

Qualitative and semi-quantitative data

Many clinical EQA schemes also evaluate the results of analytes that are not reported on a continuous scale. These may include, for example, the absence or presence of a particular pathogen species or (drug) substance and only two answers are possible: pathogen/substance present or absent. An answer that can only have two values is called dichotomous, or binary. The results of other parameters may be expressed by semi-quantitative measure, such as integer values on which arithmetic operations should be handled with caution. Traditional measures of laboratory performance, like Z- or Q-scores cannot be applied here and laboratory performance for one parameter, one sample are often limited to reporting whether the laboratory has given the consensus or expected answer or not. Although it is, for the patient's safety, extremely important to follow up individual answers for qualitative parameters that are out of consensus, like for example blood groups, combining results and counting the frequency of correct and false results for multiple samples and/or laboratories may yield additional information to evaluate analytical methods or laboratories.

For evaluating positive samples, sensitivity and positive predictive value can be used. Sensitivity is the probability of finding a positive answer for a positive sample; positive predictive value is the probability that a sample is positive when the answer is positive. Specificity is the probability of finding a negative answer for a negative sample; negative predictive value is the probability that a sample is negative if the answer is negative. Specificity and sensitivity are usually used to describe method performance, while positive and negative predictive values are more important from a clinical point of view. A combined score is the reliability, which reflects the percentage of correct results, taking into account a set of positive and negative samples. Standard errors and confidence intervals for these parameters can be calculated using standard formulas that are derived from the binomial distribution (38–40).

Similar to the usual measures of repeatability and reproducibility, new measures have been introduced (38): accordance for within laboratory agreement and concordance for between laboratory agreements. As the equivalent of repeatability, accordance reflects the probability that two identical test materials assessed by the same laboratory under standard repeatability conditions give the same result. As the equivalent of reproducibility, concordance reflects the probability that two identical test materials analysed under different conditions will give the same result. Accordance and concordance can be compared with each other to estimate the proportion of between-laboratory variation: if the concordance is smaller than the accordance, between-laboratory variation is important. Because the magnitude of concordance and accordance depends on the sensitivity, the concordance odds ratio has been introduced:

$$\text{COR} = \frac{\text{accordance } (100 - \text{concordance})}{\text{concordance } (100 - \text{accordance})}$$

where accordance and concordance are expressed as percentages (38).

Where dichotomous answers are given for a parameter that has an underlying continuous character, for example simple tests that reflect whether a substance is below or above a certain threshold, like human chorionic gonadotropin (hCG) in urine, specific EQAs can be set up with sample concentrations around the decision limit. Models have been developed to obtain estimators of central location and variability to evaluate different measurement methods (41–43). When titers are involved, the result may be dichotomized, for example by evaluating whether the reported titer would or would not lead to an incorrect conclusion (9).

Other systems to deal with qualitative tests are credit-scoring systems. Depending on the answers and their clinical impact, credit points are given or subtracted in order to obtain a final mark for the laboratory (9).

Graphical presentation for one parameter, one sample

The evaluation of laboratories and methods is greatly supported by a graphical representation of the data and is also required by international standards (8,13). To give an informative and concise summary, graphical representations should be informative with as few lines, shapes or colours as possible. Specifically for EQA, it is important to note that the graphs should not be influenced by a small fraction of heavily deviating results. There are two different types of graphs that enable laboratories to evaluate themselves with respect to their peer group or to all the participants: box plots and histograms.

Box plots are based on three different percentiles: the 25th (P25), the 50th (which is equivalent to the median) and the 75th (P75). A rectangle is drawn from P25 to the P75 percentile and lines extend the rectangle as far as values are not outliers. The outlier exclusion rule is simple and it states that all values lower than P25 - 1.5 (P75 – P25) and higher than P75 + 1.5 (P75 – P25) are considered as outliers (Figure 1). Eventually, outliers can be added as separate dots on the graph. Box plots inform about the location, scale and symmetry of the different groups, and for each group individually, show the presence - or absence - of outliers (44). Box plots adapted for EQA could be created by showing a box plot of all the data next to a box plot of the method group, with an indication of the individual laboratory result. Coloured or shaded rectangles can be used to indicate the area of acceptance according to different scoring systems. Box plots have the advantage of keeping their visual power even when they are reduced to small size and hence, they are ideal candidates for putting in reports containing results for multiple parameters.

A histogram is a classical nonparametric estimator of the distribution of the data and is today still an important statistical tool for displaying and summarizing data. Its creation is straightforward: (a) divide the interval of the data in subintervals of equal width; (b) count the number of data in each subinterval; (c) display the counts in a bar graph of which the bar heights for each subinterval correspond to the number of data in the corresponding subinterval. Histograms inform about the centre of the distribution, the possible existence of modes and the symmetry of the distribution.

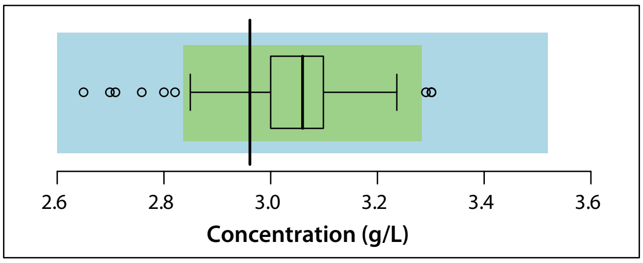

Figure 1. Box plot for evaluating an individual result
The black rectangle reaches from the 25th to the 75th percentile; the vertical line inside the black rectangle is the median. The horizontal lines to the left and the right of the box plot ('whiskers') reach to the furthest values that are closer than 1.5 times the interquartile range from the 25th and 75th percentile. The blue rectangle reflects the limits for Q-scores, the green rectangle the Z-score limits. The bold vertical line represents the individual result of a laboratory. It has a good performance according to both limits.

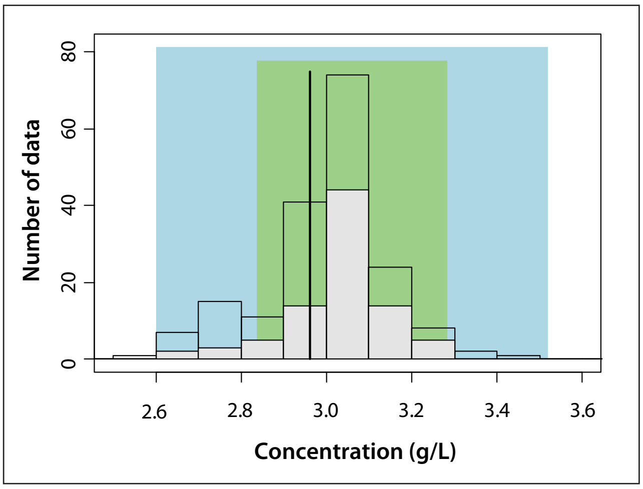

Figure 2. Histogram for the same individual result as for Figure 1. The highest bars represent all the data, the light grey bars represent the data of the peer group of the laboratory under interest. The bold vertical line represents the individual result of a laboratory. The blue rectangle reflects the limits for Q-scores, the green rectangle the Z-score limits.

The width, and consequently, number of intervals is however arbitrary. Many small subintervals lead to an irregular shaped histogram, while large and few subintervals lead to a very rough estimation of the data. Algorithms that calculate optimal subinterval widths should be applied (45).

A histogram can be easily adopted to show important information related to EQA, as illustrated in Figure 2. In case of peer group evaluation, two histograms could be superposed: the histogram of all the data, and a histogram of the peer group of the laboratory.

Evaluation intervals can be drawn by means of rectangles that are put on the background of the histogram. In this way, it is easy to estimate the fraction of data that are outside of the limits, how the own method performs with respect to the whole group and importantly, how the individual laboratory result is situated with respect to the own method group, to all the data and to the decision limits.

Graphical presentation for one parameter, multiple samples

Combining information of multiple samples can be easily done by means of a scatter plot in which the results of the laboratory are plotted against the assigned values. A robust linear regression line drawn through the points on the scatter plot not only gives a visual appraisal of the laboratory's bias but may also help the interpretation of the analytical variability or even help identifying gross outliers of which the cause may be outside the analytical phase (46).

Combining the results of two samples in a scatter plot, in which the reported results from one sample of all the laboratories are plotted against those from another, similar sample is called a Youden plot (Figure 3). Youden plots can be made of the original values or rescaled values, such as Z-scores (13,47). Some important recent developments are the addition of a robust confidence ellipse for each method (48,49). The position of the robust confidence ellipses with respect to each other reveals inter-method biases of which the interpretation is relevant for commutable samples. The position of points reflecting the values reported by individual laboratories inform about laboratory-specific bias or variability.

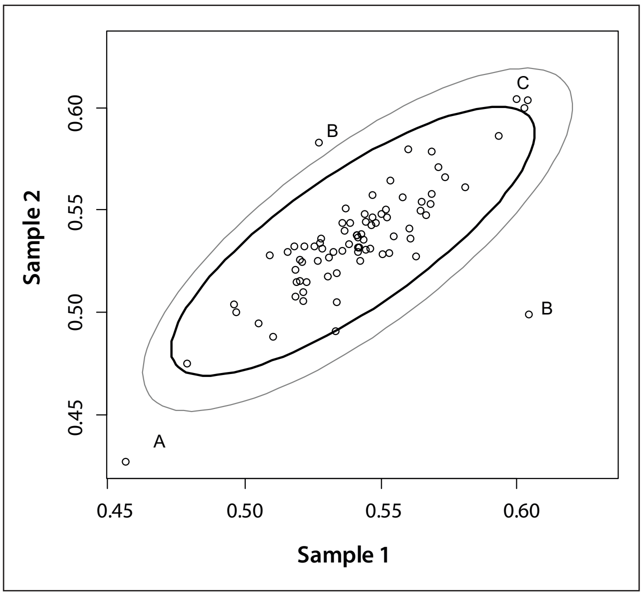

Figure 3. A Youden plot based on reported values for one specific method
The thick black line represents a 99% robust confidence region, the thin grey line a 99% confidence region based on classical statistics of average and variance-covariance matrix. Points in zone A have a negative bias, while points in zone C have a positive bias. Points in zone B exhibit high intra-laboratory variability.

Combining information from different parameters and/or samples

Several authors advised that reports could go beyond the evaluation of a certain parameter for a given sample. Combining information of multiple parameters, or multiple samples, informs about a global quality level of the laboratory and, in case samples were analysed at different time points, informs about the evolution of the quality level of the laboratory.

Results can be combined in different ways. In the first instance, laboratories might be asked to analyse the sample multiple times, in order to assess the repeatability (11). It should be noted however that two observations lead to a very uncertain measure of repeatability, and moreover, multiple analyses should always be handled with caution except when the laboratories analysed vials that have the same content but different labels (6).

In the second instance, some parameters should be considered together because the result of one parameter depends on the result of another parameter - in statistical terms: the parameters are dependent on each other. Examples are profile data, like serum electrophoresis profile or leukocyte differential count. The sum of different parameters within these profiles is a fixed value, for example, 100% in the case that the parameters represent fractions of different types that are expressed as a percentage. In this case, fractions have to be viewed as a whole. In such cases, a multivariate statistical approach is more appropriate to analyse and interpret these data. Individual laboratory evaluation is based on the multivariate distance of the laboratory results for several parameters from the centre that is made up by the assigned values of each of the parameters. This distance, the so-called Mahalanobis distance, is obtained by robust estimates of multivariate centre and variability (50). Performance characterisation of analytical methods for profile data is also possible by means of a multivariate CV, which encompasses the variability estimates of the different parameters that the profile is made of (51).

In the third instance, Z-scores can be combined in various ways. Because of their standardization with respect to the standard deviation, Z-scores are a more ideal candidate to be combined for different parameters than original reported values or Q-scores (6). A simple way to combine Z-scores is to sum them over different analytes determined for the same sample (6). Sums can be taken of (i) the Z-scores themselves (SZ); (ii) rescaling of the summed Z-scores by dividing SZ by the square root of the number of data involved (RSZ); (iii) their absolute value (SAZ) or (iv) their squared value (SSZ). Although the sum of the absolute value and the squared value leads to similar conclusions, the sum of the squared values is preferred because it has better statistical properties. It should be noted that, for a judicious interpretation of these sums, heavily deviating Z-scores often find their cause outside of the analytical process and, for this reason, they should be identified by means of an outlier test and be omitted from the calculation of the sums. If outliers are omitted, an extreme RSZ value is an indicator of bias and an extreme SAZ value is an indicator of high imprecision. Extreme values can be identified by comparing RSZ values with the standard normal distribution and SSZ values with a chi-square distribution.

Z-scores for different samples analysed over a certain period can be combined as well, some authors speak in this case of running scores (8). It is noteworthy stating that a problem from a specific round may have a ‘memory’ effect for future running scores. In this case, running scores can be smoothed by taking weighted sums of Z-scores, in a way that the influence of Z-scores on the running statistic is bigger for recent than for older Z-scores (6).

Whenever the normal distribution of the data around the assigned value cannot be assured, even not after a transformation or omitting outliers, combining Z-scores becomes cumbersome and a nonparametric approach can help evaluating laboratories by involving the reported value for multiple samples. When the difference between an individual value and the assigned value of a certain parameter for a certain sample is considered, laboratories can be ranked according to absolute value of this difference. Each reported value is allocated its own percentile value, i.e. the percentage of laboratories performing equal or worse. Subsequently, median percentile values obtained for a certain laboratory for different samples are taken and a score on a scale from 0 to 100 is obtained. Lower values indicate good performance, higher values point to weak performance (52).

Finally, results obtained for the same laboratory and parameter for samples with different assigned values can be combined by means of a linear regression model in which the independent variable is the assigned value and the dependent variable is the value found by the laboratory. Several statistics can be derived from this approach, such as the long-term coefficient of variation (LCVa) (53). It is equivalent to the variability of the points around the regression line divided by the assigned value or the long-term bias. Another statistic is the long-term bias (LTB), which is determined by the difference between the regression line and the 45-degree line reflecting equality between the assigned value and reported values. Combination of both long-term coefficients of variation and bias leads to an estimate of the uncertainty of measurement (MU) (54). It should be noted that these parameters depend largely on the assumptions of the regression model and can only be interpreted in absence of outliers and a strict linear relationship between the assigned value and reported values. In addition, the MU assumes that bias and variability are independent (54).

Another approach to the linear regression problem is first to exclude outliers from the regression model, then consider the variability of the regression model as a measure for long-term analytical variability and subsequently the bias of the regression line, after omitting regression lines with high variability (46).

Discussion

Evaluation methods applied for data gathered in EQA rounds vary widely, not only for continuous data, but also for semi-quantitative and qualitative data. For the qualitative and semi-quantitative data, it is of larger interest to combine results of different samples or surveys to estimate laboratory or method performance.

For quantitative parameters, several methods are proposed to find a consensus value or to estimate the variability. Unfortunately, there is no best method to find an assigned value or standard deviation that works well in all conditions. Although several authors attempted to compare different methods, the set of methods that were compared or the data on which they were compared varied too much to draw unique conclusions. Different methods to be used can be compared by each EQA provider using retrospective analysis on its own dataset and by means of statistical techniques that are able to estimate the uncertainty of statistical parameters with unknown distribution, like nonparametric bootstrapping (55). An alternative method is Monte Carlo simulation, a name given to any approach that uses generation of random numbers in order to find answers to specific questions. It is based on the principle that any process could be split in a series of simpler events, each presented by a probability distribution (2). The method has been applied in various studies for evaluating techniques for determining the assigned value (2,25,56) or scoring laboratories (25,57). Irrespective of the performance of each statistical method, it should not be forgotten that EQA providers have to be able to explain their statistical methods to non-statisticians in the participating laboratories. For this reason, EQA providers may prefer to use a less performing, but easy to explain statistical technique that is still able to handle outlying values.

Although combining results for different analytes or samples may reveal novel information from the reported results, it should be noted that non-experts might misinterpret scores of summed Z-values. Their general use should be handled with caution (6,8).

An important question that has not been assessed that often is the minimum number of data needed for obtaining reliable statistics. It has been mentioned that a minimum number of 20 values is necessary to have reliable robust estimates (31), although some estimators still estimate Z-scores correctly even for groups as small as 6 (25). Other authors suggest modifying the limits for evaluation of Z-scores dependent on the peer group size (50).

In conclusion, there should be no doubt that feedback reports from EQA providers to participating laboratories serve as a major tool to support their pedagogic role. Although there are mistakes that can only been detected by EQA, it should be realised however that EQA is only one aspect of the entire quality management system in laboratories. Every action undertaken based on EQA reports may be too late already. Results that were subject to the same mistake as the faulty EQA result may have been produced and reported before it could be detected by means of the EQA report. For this reason, laboratories need to reassure and implement all possible quality standards in the total testing process, since EQA reports can only serve as a follow-up of such performance (3).

Acknowledgements

The authors want to express their gratefulness to Jelena Vlasic Tanaskovic for proofreading the manuscript.

Potential conflict of interest

None declared.

References

1. Hill P, Uldall A, Wilding P. Fundamentals for external quality assessment (EQA). Available at: http://www.ifcc.org/ifccfiles/docs/fundamentals-for-eqa.pdf. Accessed February 12th 2016.
2. Wong SK. Evaluation of the use of consensus values in proficiency testing programmes. Accreditation Qual Assur 2005;10:409–14. http://dx.doi.org/10.1007/s00769-005-0029-0.
3. Visser RG. Interpretation of interlaboratory comparison results to evaluate laboratory proficiency. Accreditation Qual Assur 2006;10:521–6. http://dx.doi.org/10.1007/s00769-005-0051-2.
4. Ceriotti F. The role of External Quality Assessment Schemes in Monitoring and Improving the Standardization Process. Clin Chimica Acta 2014;432:77–81. http://dx.doi.org/10.1016/j.cca.2013.12.032.
5. Rosario P, Martínez JL, Silván JM. Comparison of different statistical methods for evaluation of proficiency test data. Accreditation Qual Assur 2008;13:493–9. http://dx.doi.org/10.1007/s00769-008-0413-7.
6. Analytical Methods Committee. Proficiency testing of analytical laboratories: organization and statistical assessment. Analyst 1992;117:97–104. http://dx.doi.org/10.1039/an9921700097.
7. Sciacovelli L, Secchiero S, Zardo L, Plebani M. External Quality Assessment Schemes: need for recognised requirements. Clin Chim Acta 2001;309:183–99. http://dx.doi.org/10.1016/S0009-8981(01)00521-6.
8. Thompson M, Wood R. The international harmonized protocol for the proficiency testing of (chemical) analytical laboratories (Technical Report). Pure Appl Chem 2009;65:2123–44.
9. Deom A, El Aouad R, Heuck CC, Kumari S, Lewis SM, Uldall A, et al. Requirements and guidance for external quality assessment schemes for health laboratories. Available at: http://www.who.int/iris/handle/10665/66089. Accessed February 3rd 2016.
10. Miller WG. Specimen materials, target values and commutability for external quality assessment (proficiency testing) schemes. Clin Chim Acta 2003;327:25–37. http://dx.doi.org/10.1016/S0009-8981(02)00370-4.
11. Miller WG, Jones GR, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. Clin Chem 2011;57:1670–80. http://dx.doi.org/10.1373/clinchem.2011.168641.
12. International Standardization Organization. 15189:2012 - Medical laboratories -- Requirements for quality and competence. Geneva, Switzerland, ISO; 2012.
13. International Standardization Organization. ISO 13528: Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparisons. Geneva, Switzerland, ISO; 2005.
14. Tholen DW. Statistical treatment of proficiency testing data. Accreditation Qual Assur 1998;3:362–6. http://dx.doi.org/10.1007/s007690050262.
15. De Bièvre P. Fitness for purpose is different from a performance specification. Accreditation Qual Assur 2007;12:501. http://dx.doi.org/10.1007/s00769-007-0312-3.
16. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, et al. Defining analytical performance specifications: Consensus Statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. Clin Chem Lab Med 2015;53:833–5. http://dx.doi.org/10.1515/cclm-2015-0067.
17. Sarmanho GF, Borges PP, Fraga ICS, Leal LH da C. Treatment of bimodality in proficiency test of pH in bioethanol matrix. Accreditation Qual Assur 2015;20:179–87. http://dx.doi.org/10.1007/s00769-015-1133-4.
18. Mancin M, Toson M, Grimaldi M, Barco L, Trevisan R, Carnieletto P, et al. Application of bootstrap method to evaluate bimodal data: an example of food microbiology proficiency test for sulfite-reducing anaerobes. Accreditation Qual Assur 2015;20:255–66. http://dx.doi.org/10.1007/s00769-015-1141-4.
19. Ellison SLR. Performance of MM-estimators on multi-modal data shows potential for improvements in consensus value estimation. Accreditation Qual Assur 2009;14:411–9. http://dx.doi.org/10.1007/s00769-009-0571-2.
20. Lowthian PJ, Thompson M. Bump-hunting for the proficiency tester—searching for multimodality. Analyst 2002; 127:1359–64. http://dx.doi.org/10.1039/B205600N.
21. Barnett V, Lewis T, eds. Outliers in Statistical Data, 3rd ed. Chichester, UK: John Wiley & sons, 1994.
22. Davies PL. Statistical evaluation of interlaboratory tests. Fresenius Z Für Anal Chem 1988;331:513–9. http://dx.doi.org/10.1007/BF00467041.
23. Kandler W, Schuhmacher R, Roch S, Schubert-Ullrich P, Krska R. Evaluation of the long-term performance of water-analyzing laboratories. Accreditation Qual Assur 2003;9:82–9. http://dx.doi.org/10.1007/s00769-003-0696-7.
24. Hund E, Massart DL, Smeyers-Verbeke J. Inter-laboratory studies in analytical chemistry. Anal Chim Acta 2000;423:145–65. http://dx.doi.org/10.1016/S0003-2670(00)01115-6.
25. Coucke W, China B, Delattre I, Lenga Y, Van Blerk M, Van Campenhout C, et al. Comparison of different approaches to evaluate External Quality Assessment Data. Clin Chim Acta 2012;413:582–6. http://dx.doi.org/10.1016/j.cca.2011.11.030.
26. Huber P. Robust statistics. New York, NY: John Wiley and sons, 1981. http://dx.doi.org/10.1002/0471725250.
27. Kandler W, Schuhmacher R, Roch S, Schubert-Ullrich P, Krska R. Evaluation of the long-term performance of water-analyzing laboratories. Accreditation Qual Assur 2003;9:82–9. http://dx.doi.org/10.1007/s00769-003-0696-7.
28. Duewer DL. A comparison of location estimators for interlaboratory data contaminated with value and uncertainty outliers. Accreditation Qual Assur 2008;13:193–216. http://dx.doi.org/10.1007/s00769-008-0360-3.
29. Ellison SLR. Performance of MM-estimators on multi-modal data shows potential for improvements in consensus value estimation. Accreditation Qual Assur 2009;14:411–9. http://dx.doi.org/10.1007/s00769-009-0571-2.
30. Yohai VJ. High breakdown-point and high efficiency robust estimates for regression. Ann Stat 1987;642–656. http://dx.doi.org/10.1214/aos/1176350366.

31. Tholen DW. Statistical treatment of proficiency testing data. Accreditation Qual Assur 1998;3:362–6. http://dx.doi.org/10.1007/s007690050262.
32. Coucke W, Charlier C, Lambert W, Martens F, Neels H, Tytgat J, et al. Application of the Characteristic Function to Evaluate and Compare Analytical Variability in an External Quality Assessment Scheme for Serum Ethanol. Clin Chem 2015;61:948–54. http://dx.doi.org/10.1373/clinchem.2015.240176.
33. Thompson M. The Characteristic Function, a Method-Specific Alternative to the Horwitz Function. J AOAC Int. 2012;95:1803–6. http://dx.doi.org/10.5740/jaoacint.12-042.
34. Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. J Am Stat Assoc 1993;88:1273–83. http://dx.doi.org/10.1080/01621459.1993.10476408.
35. Wilrich P-T. Robust estimates of the theoretical standard deviation to be used in interlaboratory precision experiments. Accreditation Qual Assur 2007;12:231–40. http://dx.doi.org/10.1007/s00769-006-0240-7.
36. Coucke W, China B, Delattre I, Lenga Y, Van Blerk M, Van Campenhout C, et al. Comparison of different approaches to evaluate External Quality Assessment Data. Clin Chim Acta 2012;413:582–6. http://dx.doi.org/10.1016/j.cca.2011.11.030.
37. Tukey JW. Exploratory data analysis. Reading, MA: Addison-Wesley, 1977.
38. Langton SD, Chevennement R, Nagelkerke N, Lombard B. Analysing collaborative trials for qualitative microbiological methods: accordance and concordance. Int J Food Microbiol 2002;79:175–81. http://dx.doi.org/10.1016/S0168-1605(02)00107-1.
39. Ellison SLR, Fearn T. Characterising the performance of qualitative analytical methods: Statistics and terminology. TrAC Trends Anal Chem 2005;24:468–76. http://dx.doi.org/10.1016/j.trac.2005.03.007.
40. Cárdenas S, Valcárcel M. Analytical features in qualitative analysis. TrAC Trends Anal Chem 2005;24:477–87. http://dx.doi.org/10.1016/j.trac.2005.03.006.
41. Petersen PH, Christensen NG, Sandberg S, Nordin G, Pedersen M, Nordic Control Organizations. How to deal with semi-quantitative tests? Application of an ordinal scale model to measurements of urine glucose. Scand J Clin Lab Invest. 2009;69:662–72. http://dx.doi.org/10.3109/00365510902968756.
42. Petersen PH, Christensen NG, Sandberg S, Nordin G, Pedersen M, Nordic Control Organizations. How to deal with semi-quantitative tests? Application of an ordinal scale model to measurements of urine glucose. Scand J Clin Lab Invest 2009;69:662–72. http://dx.doi.org/10.3109/00365510902968756.
43. Bjerner J. Comment on the article entitled "How to Deal with Dichotomous Tests? Application of a Rankit Ordinal Scale Model with examples from the Nordic Ordinal Scale Project on screening tests" by Petersen et al. Scand J Clin Lab Invest 2009;69:28–30. http://dx.doi.org/10.1080/00365510802290681.
44. Williamson DF, Parker RA, Kendrick JS. The Box Plot: A Simple Visual Method to Interpret Data. Ann Intern Med 1989;110:916–21. http://dx.doi.org/10.7326/0003-4819-110-11-916.
45. Scott DW. On optimal and data-based histograms. Biometrika 1979;66:605–610. http://dx.doi.org/10.1093/biomet/66.3.605.
46. Coucke W, Van Blerk M, Libeer J-C, Van Campenhout C, Albert A. A new statistical method for evaluating long-term analytical performance of laboratories applied to an external quality assessment scheme for flow cytometry. Clin Chem Lab Med 2010;48:645–650. http://dx.doi.org/10.1515/CCLM.2010.122.
47. Youden WJ. Graphical diagnosis of interlaboratory test results. Stat Concepts Proced. 1969;1:133.
48. Shirono K, Iwase K, Okazaki H, Yamazawa M, Shikakume K, Fukumoto N, et al. A study on the utilization of the Youden plot to evaluate proficiency test results. Accreditation Qual Assur 2013;18:161–74. http://dx.doi.org/10.1007/s00769-013-0978-7.
49. Zhou Q, Hu J, Li X, Li S, Gao Z, Xie W, et al. Comparison of traditional, trimmed traditional and robust Youden charts. Clin Chim Acta 2015;446:213–7. http://dx.doi.org/10.1016/j.cca.2015.04.021.
50. Zhang L, Campenhout CV, Devleeschouwer N, Libeer J-C, Albert A. Statistical analysis of serum protein electrophoresis results in External Quality Assessment schemes. Accreditation Qual Assur 2008;13:149–55. http://dx.doi.org/10.1007/s00769-008-0388-4.
51. Zhang L, Albarède S, Dumont G, Campenhout CV, Libeer J-C, Albert A. The multivariate coefficient of variation for comparing serum protein electrophoresis techniques in external quality assessment schemes. Accreditation Qual Assur 2010;15:351–7. http://dx.doi.org/10.1007/s00769-009-0627-3.
52. Ehrmeyer SS, Laessig RH. Alternative statistical approach to evaluating interlaboratory performance. Clin Chem 1985;31:106–8.
53. Meijer P, de Maat MPM, Kluft C, Haverkate F, van Houwelingen HC. Long-Term Analytical Performance of Hemostasis Field Methods as Assessed by Evaluation of the Results of an External Quality Assessment Program for Antithrombin. Clin Chem 2002;48:1011–5.
54. Matar G, Poggi B, Meley R, Bon C, Chardon L, Chikh K, et al. Uncertainty in measurement for 43 biochemistry, immunoassay, and hemostasis routine analytes evaluated by a method using only external quality assessment data. Clin Chem Lab Med 2015;53:1725–36. http://dx.doi.org/10.1515/cclm-2014-0942.
55. Thompson M. The variance of a consensus. Accreditation Qual Assur 2006;10:574–5. http://dx.doi.org/10.1007/s00769-005-0037-0.
56. Duewer DL. A comparison of location estimators for interlaboratory data contaminated with value and uncertainty outliers. Accreditation Qual Assur 2008;13:193–216. http://dx.doi.org/10.1007/s00769-008-0360-3.
57. Wong S. Performance evaluation for proficiency testing with a limited number of participants. Accreditation Qual Assur 2011;16:539–44. http://dx.doi.org/10.1007/s00769-011-0816-8.