Opinion Paper

Graham Ross Dallas Jones*

Analytical performance specifications for EQA schemes – need for harmonisation

DOI 10.1515/cclm-2014-1268
Received December 22, 2014; accepted March 18, 2015; previously published online April 17, 2015

Abstract: External Quality Assurance (EQA) is a vital tool in laboratory medicine to assess individual laboratory analytical performance and also the differences between the results from different laboratories. This information is also useful for professional bodies and manufacturers as part of post-market surveillance. The process involves the measurement of one or more samples by many laboratories and then assessment of the results. Individual results are generally assessed by how far they lie from a target, which may be established using reference methods or a median of some or all of the submitted results. The distance of a result from the target is compared with analytical performance specifications in order to assess the analytical quality. One of the uses of the Stockholm hierarchy of performance goals is to set the performance specifications for analysis of EQA results. Fifteen years after the Stockholm consensus meeting, EQA analytical performance specifications appear to still vary widely between EQA providers. This can be due to a range of factors, including the rationale for setting the criteria, the expected response to a failure to meet the specified performance, the clinical meaning behind meeting the specifications, and the possible need for further analytical improvements. There are also differences in the models chosen to set the criteria, usually either state of the art or biological variation, and then differences in how these are applied. While harmonisation of EQA performance specifications may be some time off, all EQA providers should define the nature of their specifications and the basis for their selection and make this information available to customers.

***Corresponding author: Graham Ross Dallas Jones,** Department of Chemical Pathology, St. Vincent's Hospital, Sydney, 390 Victoria Street, Darlinghurst, NSW 2010, Australia; and Faculty of Medicine, University of NSW, Kensington, NSW, Australia,
Phone: +61 2 83829160, E-mail: Graham.Jones@svhm.org.au

Keywords: analytical performance criteria; External Quality Assurance; proficiency testing.

Introduction

The primary role of External Quality Assurance (EQA) for laboratories is to confirm assay performance when it is performing well, to identify poor assay performance when it occurs, and then confirm correction of poor performance [1]. The main issues that are assessed by EQA programs are accuracy (meaning the uncertainty caused by the combination of bias and precision), and bias and precision separately. Additionally, EQA may assess analytical specificity, interferences, units, reference intervals, calculations [2], and interpretation [3]; however, these are not the subjects of this paper. EQA is particularly relevant as part of a discussion about analytical performance specifications as it is a place where quality standards are applied, and, in this setting, they can be used for assessment of the combined effect of all other analytical quality activities.

In brief, the usual process of performing EQA is the following sequence: the EQA provider prepares and distributes samples; laboratories analyse the samples and submit results back to the provider; and the provider then prepares a report that is received and interpreted by the participating laboratories. Any performance specifications are applied to the data on the report, and the effect of the specifications is in the actions taken by laboratories in response to the reports. More details concerning the basic components of EQA are available in ISO 17043 [4]. It is worth also considering that in addition to individual laboratories, EQA data can also be reviewed by professional organisations, manufacturers, health-care providers, and health funding bodies. Manufacturers can use EQA results for post-marketing surveillance to alert themselves to any analytical problems as described in the standard EN 14136 [5], and the wider pathology community may wish to use the data to consider whether the analytical performance from participating laboratories would indicate that common reference intervals and decision points are

appropriate. EQA programs from different providers, even if they are for the same analytes, differ from each other in many ways. There are many EQA providers from different locations in the world that provide high-quality, robust, and innovative programs; however, for the purposes of this discussion, I will use terminology and examples from the Royal College of Pathologists of Australasia Quality Assurance Program (RCPAQAP) in Australia, as this is the program with which I am most familiar both as a user of the program and as current Chair of the Advisory Committee of the RCPAQAP Chemical Pathology program.

EQA reporting

EQA reports come in many different formats but it is common for providers to produce a report after each challenge (a set of samples analysed at a single time), which is sometimes called an interim report [6]. Typically, there are a small number of samples per challenge (e.g., 1, 2, or 5) on which the report is based, although an interim report may also include data from previous challenges. Even if more than one sample is analysed in a challenge, the samples are often assessed separately as single results. As a single result includes the effects of both bias and imprecision, the performance specification applied assesses "total error". This also applies to multiple samples if they are analysed separately.

Reports based on a series of challenges over a period of time, which may be referred to as end-of-cycle or summary reports, generally include the results of enough samples to separately analyse bias and precision. Thus, the quality specifications in these reports can be applied to separately consider these two aspects of analytical performance. In this setting, it is important to recognise that more data allows for improved estimates of bias and precision. From the above, it can be seen that separate standards are required for analysing single results and for analysing bias and precision based on multiple results; however, the focus of this paper is the analytical quality standards for interpreting single results.

Analytical performance specifications

An EQA report based on a single result includes at least the following information: the result from laboratory and a target from the EQA program, which together define the distance of the result from the target. There is then an assessment of this distance from the target, which can be qualitative (inside or outside the limit) or quantitative, indicating the extent of the distance beyond the limit. In the same way that all interpretations of numerical pathology results is based on comparison [7], the deviation of the result from the target is assessed by comparison with a quality specification.

A key component of the interpretation is the selection of the target to indicate the "correct" result for the comparison. There are two main types of target for EQA programs: an overall target for all results or method-specific targets for a subset of results. Overall targets may be based on a reference method or material or an overall result median. The optimal value of an overall target is achieved when material with verified commutability in the methods used in the program is used in combination with value assignment with a reference measurement procedure [1]. Alternatively, programs may supply method-specific targets based on method, instrument, or reagents. The choice of target is vital for quality assessment in EQA programmes; however, is not the topic for this presentation.

There is no currently available data on quality standards used in EQA programs; however, a study from 1996 showed very wide ranges [8]. For example, the limits from cholesterol range between 3% and 18%, for phosphate between 5% and 14%, and for alkaline phosphatase between 7% and 30%. Other current sources such as the RCPAQAP, the Clinical Laboratory Improvements Act (CLIA) [9] in the US, and RiliBÄK in Germany indicate that performance specifications remain very different [10]. These differences may be due to a range of factors, including the rationale for setting the criteria, the expected response to a failure to meet the specified performance, the clinical meaning behind meeting the specifications, and the possible need for further analytical improvements. Other factors may be the clinical setting (e.g., point of care compared with main laboratory) and available economic resources in different regions.

To understand the contributors to the variation in analytical performance specifications, it is necessary to consider the rationale used for setting the limits by each EQA provider. This can be considered in several ways. The first question would be what type of specification is being supplied. This may vary along a continuum from a minimum specification, which all reasonable laboratories would be expected to pass; an expected specification, which most laboratories should pass but with the aim to improve those that do not meet the specification; through to an aspirational specification where some or many laboratories will not meet until better methods are developed.

Across this continuum, the specifications would move from looser to tighter.

The setting of analytical performance specifications can vary with respect to the expected response to results outside the specifications. Some limits are used for regulatory purposes and can affect a laboratory's ability to perform testing. An example is the CLIA regulations in the US [9]. In some settings, failures may require mandatory investigation involving time and effort with compliance; in other settings, results outside limits should be followed up, with the amount of effort depending on the nature and severity of the failure. The Australian accreditation environment could be described in this way. If aspirational limits are set beyond the performance of current methods, the response must come from industry rather than individual laboratories. Again, looser limits would be expected in environments associated with a more significant response to failures.

Analytical performance specifications can also have different implications for the assessment of the assay performance. For assays meeting a looser standard, there may still be benefits from further assay improvement, whereas an assay meeting a very tight standard may indicate that no further effort is needed for measurement of this analyte. Additionally, meeting quality specifications of a different level can lead to knowledge about the appropriate clinical use of test results. For example, if assays meet a very tight standard, this might indicate that patients can be monitored at an optimal level among the laboratories meeting this standard. At a looser standard, monitoring may be less effective but sharing of reference intervals may be supported. Meeting very loose quality standards may not even indicate that common decision points are valid and separate reference intervals are required.

The Stockholm hierarchy

One of the outcomes of the 1999 Stockholm consensus [11] may have been an expectation that the application of this approach may have brought the quality standards together. Even where the criteria have been applied, it is fair to say that, "If you have seen one implementation of the Stockholm hierarchy, you have seen one implementation of the Stockholm hierarchy".

When attempting to apply the Stockholm criteria, some inbuilt contradictions in the hierarchy become apparent. For example, level 1 of the hierarchy (based on clinical outcomes) and level 2b (based on physician opinion) are both dependent on current analytical performance (level 5). This is also the case with professional recommendations (level 3), as the experts making the recommendations are aware of the state of the art. Additionally, if a professional organisation (level 3) or EQA organiser (level 4) uses an approach based on level 1 or 2, does this then return to level 3 when they make the recommendations or do their recommendations a step higher due to the approach used?

The revised structure being proposed for consideration at this meeting allows the selection from one of three models as follows: model 1 – based on the effect of analytical performance on clinical outcomes; model 2 – based on components of biological variation of the measurand; model 3 – based on state of the art. This new proposal removes some of these inconsistencies, as quality specifications defined by the organisation that established the limits are no longer considered as bases for a category. The concept that current analytical performance affects the other levels, however, remains true.

It is also possible to provide multiple specifications on a single report. These may be multiple levels of the same type of standard, e.g., analytical performance reported against the optimal, desirable, or minimal levels based on biological variation. There may also be different types of standards, e.g., statistical and clinically based standards, on the same report. Of course, then, it is possible that a result may meet one standard and fail another. This use of dual standards for the state of the art and total error based on biological variation is practised by SKML (Dutch Foundation for Quality Assessment in Medical Laboratories) in the Netherlands [12].

As stated above, the process of applying the Stockholm criteria is done by people, usually as part of organisations. The inputs to the process would be selection of a background principle, seeking information on clinical studies, biological variation (e.g., the Ricos database [13]), and current analytical performance (e.g., EQA data). It is my contention that even given the same data and the same conceptual approach, laboratory scientists will interpret the data differently and arrive at different conclusions. Thus, different performance specifications in different EQA programs are an expected outcome unless specific steps are taken to seek uniformity. This can be seen in the setting of reference intervals by laboratories where variability rather than commonality among reference intervals is the usual outcome [14].

As there is a limited amount of clinical outcome data to support specific analytical quality standards, in assessing the use of the three levels of the proposed revised

hierarchy, I will concentrate on the state of the art and biological variation.

"State-of-the-art" analysis of single results is generally performed by comparing against the range of other submitted results for the same sample. This is most commonly done by statistical analysis where the target is usually an estimate of the middle of a group of results and the limits are typically ±2 or 3 standard deviations (SDs) of the group. Obviously, the number of outliers will depend on the way the limits are set. Additionally, a quantitative or severity assessment can be made by the use of z-scores or similar numerical process (comparing the difference of the submitted result with the scatter of submitted result differences). Under the 1999 hierarchy, this approach would be described as level 5, state of the art, and as model 3 in the recent proposal.

This statistical analysis compares a laboratory with other similar laboratories and can alert to possible analytical or work practice problems, although the clinical meaning of a result outside statistical limits is uncertain. Within this type of approach, there are also areas of difference where commonality of approach would be required for standardisation. These factors may include outlier exclusion; use with other limits; limits set at 2SD, 3SD, or other; handling of small method groups; and processes to identify method groups.

In practice, higher-level performance specifications, e.g., levels 1–4 from the 1999 consensus, are based on biological variation. Even after accepting this level of the hierarchy, there are a wide range of possible criteria for standard selection. By way of example, I will briefly outline the processes used at the RCPAQAP and contrast with other possibilities. These have been described in more detail previously [15].

RCPAQAP allowable limits of performance

The RCPAQAP quality standards, known in the program as "allowable limits of performance" (ALP) are the analytical range around a central value (the target) that provides a simple tool to allow a rapid, standardised assessment of QAP results in both numerical and graphical report formats. A result outside the ALP should alert the laboratory that that their assay may produce results that are at risk of detrimentally affecting clinical decision-making. Of note, the limits are designed neither to be regulatory nor as an optimal standard for all assays. One description of the ALP would be that they are the "reference intervals" of EQA reports, drawing attention to possible problems but neither confirming nor excluding the "disease" of the assay.

We agreed that our limits would be used to assess total error, as they were applied to single results as described above. We also agreed to select percentage limits from a specified list (1%, 2%, 3%, 4%, 5%, 6%, 8%, 10%, 12%, 15%, 20%, 25%, and 30%) rounding to the nearest of these values as a recognition that the background data for decision-making was itself of limited precision. As is done for some other programs, the limits also include a change between absolute and percentage values based on precision profile [9].

For each analyte, a criterion was selected either based on the within-subject biological variation (referred to as imprecision criteria although applied to total error of the results), or on the combined within- and between-subject variation (referred to as total error criteria). Within these larger categories, a further selection was made from the optimal, desirable, and minimal levels as described by Fraser [16]. The selection process became a balance between a decision to not set unachievable goals as well as a stated aim to improve laboratory performance. Thus, a limit that was wide and could be met by everyone may not drive improvement and a limit that was unachievable would lead to the limits being ignored. We agreed to set the tightest limit based on biological variation would be selected within the limitations of the current state of the art, for example, the performance that could be achieved by about 80% of laboratories. The group setting the limits would gather data on current state of the art (from our own EQA data), latest biological variation data, and any other relevant professional recommendations. All the data would be recorded as well as the factors taken into account in setting the limits to make the decision "traceable" to the available data.

On the basis of these levels of criteria, the following use of the data could be confirmed if all data from multiple laboratories could meet the criterion for an analyte. If a criterion based on "total error" was met, then the laboratories could share a common reference interval for the analyte. If the "imprecision" criterion is met, a patient can be monitored successfully across the laboratories. Within these categories, "optimal" indicates no need to improve further, "desirable" indicates satisfactory performance, and "minimal" indicates room for improvement in the assay.

The ALP have been used as criteria for assessment of data for consideration of common reference intervals in Australian laboratories. When the majority of individual results from a method comparison study are within the

limits, it is valid to share intervals, and this approach is one factor in the recent setting of recommended intervals for 15 analytes [17].

Analysis at a recent Australian Quality Control workshop [18] demonstrated that the ALP for a majority of analytes could be used as the first step for quality planning process by comparing analytical precision with the limits and determining the sigma value. This was not possible for all analytes, indicating that analytical improvements are required. One Australian approach is to assess the sigma value from individual laboratories and also from the best-performing laboratories to identify whether a low sigma value can be fixed by an individual laboratory, or whether an industry approach is required [19].

Conclusions

So are we ready for harmonised EQA quality performance specifications? In short, no, or at least not yet. This can only happen with a significant collaborative effort with clearly defined and agreed goals. At this time, the purposes for which EQA performance criteria are used are widely varied, and so differences in the specifications themselves is expected. As a starting point, all EQA programs should provide their customers at least the following information about their quality specifications: the nature of the specifications (minimal, aimed at driving improvement, optimal), the expected response to results outside the limits, how the limits were determined, and what the effect of compliance means for interpretation of the results. EQA programs may also choose to provide more than one type of standard or more than one level of standard of the same type.

While an international harmonising EQA quality standards is unlikely, the use of common terminology and agreed approaches to setting limits may allow progress towards sharing of data and improvements where they are most needed.

Acknowledgments: I acknowledge the major inputs of Ken Sikaris and Janice Gill into the work of the RCPAQAP in setting the allowable limits of performance.
Author contributions: The author has accepted responsibility for the entire content of this submitted manuscript and approved submission.
Financial support: None declared.
Employment or leadership: None declared.
Honorarium: None declared.

References

1. Miller WG, Jones GRD, Horowitz GL, Weykamp C. Proficiency testing/external quality assessment: current challenges and future directions. Clin Chem 2011;57:1670–80.
2. Jones GRD, Koetsier S. RCPAQAP chemical pathology calculated results program – first report. Presented at RCPA Update, Melbourne, February 2014 (poster).
3. Lim EM, Sikaris KA, Gill J, Calleja J, Hickman PE, Beilby J, et al. Quality assessment of interpretative commenting in clinical chemistry. Clin Chem 2004;50:632–7.
4. ISO/IEC 17043. Conformity assessment – general requirements for proficiency testing. International Standards Organisation, 2010.
5. EN14136. Use of external quality assessment schemes in the assessment of the performance of in vitro diagnostic procedures. European Standards, 2004.
6. Reports. RCPAQAP. Available at: www.rcpaqap.com.au. Accessed: 11 Dec 2014.
7. Petersen PH. Making the most of a patient's laboratory data: optimisation of signal-to-noise ratio. Clin Biochem Rev 2005;26:91–6.
8. Ricos C, Baadenhuijsen H, Libeer CJ, Petersen PH, Stockl D, Thienpont L, et al. External quality assessment: currently used criteria for evaluating performance in European countries, and criteria for future harmonization. Eur J Clin Chem Clin Biochem 1996;34:159–65.
9. Clinical Laboratory Improvements Act. Title 42 – public health, Chapter IV – Centers for Medicare & Medicaid Services, Department of Health and Human Services. Subchapter G – standards and certification, Part 493 – laboratory requirements, Subpart I – proficiency testing programs for nonwaived testing, Subgrp – proficiency testing programs by specialty and subspecialty. US Government Printing Office. Available at: http://www.gpo.gov/fdsys/granule/CFR-2011-title42-vol5/CFR-2011-title42-vol5-sec493-931/content-detail.html. Accessed: 14 Dec 2014.
10. RiliBak. Neufassung der "Richtlinie der Bundesärztekammer zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen – Rili-BÄK". Available at: http://www.bundesaerztekammer.de/downloads/Rili-BAEK-Laboratoriumsmedizin.pdf Accessed: 14 Dec 2014. English version: Revision of the "Guideline of the German Medical Association on quality assurance in medical laboratory examinations – Rili-BAEK." J Lab Med 2015;39:26–69.
11. Kenny D, Fraser CG, Hyltoft Petersen P, Kallner A. Strategies to set global analytical quality specifications in laboratory medicine – consensus agreement. Scand J Clin Lab Invest 1999;59:585.
12. MUSE scoring and reporting system. Dutch Foundation for Quality Assessment in Medical Laboratories (SKML). Available at: http://www.skml.nl/uploads/48/fc/48fca1596699ea73af3b74d9cb841f24/MUSE-manual-version-2.1.pdf. Accessed: 12 Dec 2014.
13. Ricos C. Desirable biological variation database specifications. Available at: https://www.westgard.com/biodatabase1.htm. Accessed: 12 Dec 2014.
14. Jones GRD, Koetsia S. Combined reference intervals and analytical EQA. Clin Biochem Rev 2014;35:243–50.

15. Jones GRD, Sikaris K, Gill J. 'Allowable limits of performance' for External Quality Assurance programs – an approach to application of the Stockholm criteria by the RCPA quality assurance programs. Clin Biochem Rev 2012;33:133–9.
16. Fraser CG. Biological variation: from principles to practice. AACC Press, Washington, DC, 2001.
17. Tate JR, Sikaris KA, Jones GRD, Yen T, Koerbin G, Ryan J, et al. Adult and paediatric common reference intervals in Australia and New Zealand for a first panel of chemistry analytes. Clin Biochem Rev 2014;35:213–35.
18. AACB Quality Control Workshop Part 2 – Harmonisation of Approaches to Setting a Laboratory QC Policy. Accessed: 30 Sep 2014, Adelaide.
19. Bais R. Use of capability index to improve laboratory analytical performance. Clin Biochem Rev 2008;29 (Suppl 1):S27–31.