Clinical Biochemistry 49 (2016) 699–707

Contents lists available at ScienceDirect

Clinical Biochemistry

journal homepage: www.elsevier.com/locate/clinbiochem

Analytical

Utilizing global data to estimate analytical performance on the Sigma scale
A global comparative analysis of methods, instruments, and manufacturers through external quality assurance and proficiency testing programs

Sten A. Westgard

Westgard QC, Inc., Madison, WI, USA

ARTICLE INFO

Article history:
Received 4 February 2016
Received in revised form 25 February 2016
Accepted 26 February 2016
Available online 4 March 2016

Keywords:
Quality
External quality assessment
Proficiency testing
Six sigma
Sigma-metrics
Chemistry
Instrument performance

ABSTRACT

Objective: To assess the analytical performance of instruments and methods through external quality assessment and proficiency testing data on the Sigma scale.

Design and methods: A representative report from five different EQA/PT programs around the world (2 US, 1 Canadian, 1 UK, and 1 Australasian) was accessed. The instrument group standard deviations were used as surrogate estimates of instrument imprecision. Performance specifications from the US CLIA proficiency testing criteria were used to establish a common quality goal. Then Sigma-metrics were calculated to grade the analytical performance.

Results: Different methods have different Sigma-metrics for each analyte reviewed. Summary Sigma-metrics estimate the percentage of the chemistry analytes that are expected to perform above Five Sigma, which is where optimized QC design can be implemented. The range of performance varies from 37% to 88%, exhibiting significant differentiation between instruments and manufacturers. Median Sigmas for the different manufacturers in three analytes (albumin, glucose, sodium) showed significant differentiation.

Conclusions: Chemistry tests are not commodities. Quality varies significantly from manufacturer to manufacturer, instrument to instrument, and method to method. The Sigma-assessments from multiple EQA/PT programs provide more insight into the performance of methods and instruments than any single program by itself. It is possible to produce a ranking of performance by manufacturer, instrument and individual method. Laboratories seeking optimal instrumentation would do well to consult this data as part of their decision-making process. To confirm that these assessments are stable and reliable, a longer term study should be conducted that examines more results over a longer time period.

1. Introduction

External Quality Assurance (EQA), also known as Proficiency Testing (PT), is considered a mandatory component of a laboratory quality system. ISO 15189 requires EQA [1]. CLIA regulations in the US require PT for all regulated and non-regulated analytes that are non-waived [2]. The purpose of EQA/PT is to provide an external check on the quality of the laboratory, particularly an assessment of the bias and accuracy of the method.

A more modern assessment of quality can be attained through the use of the Sigma-metric [3]. Six Sigma metrics combine, bias, imprecision, and the allowable total error and convert that into an overall assessment of the analytical quality of the test. The concepts of Six Sigma have been around industry and healthcare for decades. The Six Sigma scale is a well-accepted and popularly understood set of benchmarks that can be applied to any process. Any test or process that achieves the eponymous Six Sigma operates nearly defect-free, while a process or test that operates below 3 Sigma is typically considered unstable for routine operation. The amount of QC required for a Six Sigma testing method is dramatically lower than the requirements of a 3 Sigma or lower method, and the Sigma-metric can be integrated into the Risk Management performed by the laboratory [4].

The analytical Sigma-metric has been used on chemistry methods [5], enzyme methods [6], immunoassay methods [7], point-of-care instruments [8], and highly standardized methods such as HbA1c [9]. There are applications that have been described not only in the "established" laboratory world [10], but also in developing countries like Ghana [11], Egypt [12], India [13,14], etc. The Sigma-metric is the analytical assessment model recommended by the AACB [15], CLSI C24 [16], as well as the Task Force on the Implementation of HbA1c standardization (TF-HbA1c) [17].

E-mail address: westgard@westgard.com.

http://dx.doi.org/10.1016/j.clinbiochem.2016.02.013
0009-9120/

In 2006, Westgard and Westgard evaluated Proficiency Testing data from five different US PT programs using the Sigma-scale and found that analytical quality was not performing up to the necessary level [18]. In 2013, Jassam et al. used Sigma-metrics to examine the performance of UK labs and their ability to support evidence-based guidelines for diabetes and ischaemic heart disease [19] — finding that only a minority of labs were providing the necessary quality in their testing. These previous studies aggregated performance from different labs and instruments, without attempting to determine if instrument performance varied or contributed to a lack of quality. In 2015, Westgard and Westgard introduced a Proficiency Testing Sigma-Metric chart, a graphic tool which allowed the assessment of EQA/PT data on the Six Sigma scale, particularly for HbA1c and other methods that have commutable, accuracy-based programs [20]. The application of Sigma-metrics on EQA/PT data is well-established, but a global assessment across multiple programs has not yet been conducted.

Despite the near-uniform requirement of participation in EQA/PT by most of the world, the practice of EQA/PT is not standardized. Quality requirements (in the parlance of ISO metrology, analytical performance specifications) vary by program and country. The use of non-commutable samples is extremely common, but while EQA/PT "provides useful information to a single participant [it] does not allow improvements in harmonization and inter-laboratory agreement of results" [21]. In 1996, Ricos et al. [22] found that chemistry performance specifications varied as much as 200% for common analytes. In 2005, Friedecky et al. [23] reaffirmed that these disparities in performance specifications persist and progress in harmonization has been slow. More, recently Graham Jones [24] noted that the differences in performance specifications in EQA/PT programs in part derives from their different purposes. Programs that impose harsh penalties for failure, such as de-funding or closing a laboratory, tend to have more lenient performance specifications (i.e. CLIA and the German Rilibak). EQA/PT programs that are more educational in nature (that is, that do not impose severe sanctions for failures) are often more stringent in their performance specifications.

Thus, current EQA/PT programs pose a quandary for laboratories: Does your EQA success depend on the country in which your laboratory resides? Can methods in Germany and the US perform worse because there are more lenient performance specifications? Does the acceptability of method and laboratory performance also depend upon the *specific* EQA/PT program in which the laboratory participates? Will a laboratory that participates in a European program experience more EQA/PT failures than a laboratory that participates in a US survey? It seems illogical that a method can be acceptable in one part of the world, yet unacceptable in another part, yet have the same performance.

The task of standardization of EQA/PT programs is monumental. Not only do the performance specifications differ, but so do the statistics, specimens, testing frequency, and nearly every other aspect. EQA/PT programs seek to differentiate themselves from their competitors by offering unique statistical calculations and graphical displays. Thus, robust competition among EQA/PT providers often frustrates the standardization efforts. Because of this, Graham Jones recently concluded that while it may be hoped that EQA/PT programs will soon provide their results in a common terminology, it is not likely that true standardization will occur for decades.

Setting aside these challenges, the data currently collected by EQA/PT programs around the world still represent a unique resource for today's laboratories. If the data from EQA/PT programs could be compared in a standardized way, these results would offer a global perspective on laboratory, method, and instrument performance. If these large sets of data were to be mined, the 'Big Data' analytics could reveal not only what performance specifications are realistic, but also what methods, instruments, and manufacturers provide the best performance. Just as Google can sift through millions of links to find the most appropriate match to a search query, the utilization of EQA/PT data from multiple programs can aid laboratories find the most appropriate method for their clinical needs.

2. Materials and methods

Data from five different EQA/PT programs were obtained, which, out of sensitivity to the EQA/PT programs, will be kept anonymous. First, from an American EQA program that has significant global participation (hereafter A), followed by an American EQA program with a mostly domestic participation (hereafter B), a Canadian EQA program (hereafter C), a United Kingdom-based EQA program (hereafter D), and finally an EQA program based in Australia with significant participation throughout Asia (hereafter E).

As stated earlier, the data gathered and analysis provided by each EQA/PT program is not standardized. In some reports, a consensus mean was calculated for all methods. In other cases, only principle-specific means were calculated, and no overall group mean was determined. This of course complicates any attempt to determine findings across the programs.

Specimens from program A are provided in five 5.0-mL liquid serum specimens, three shipments per year. Specimens from program B are provided in five 5 mL serum specimens, three times per year. Specimens from chemistry program C are provided in three 1.5 mL previously frozen pooled human serum specimens supplemented with selected analytes, three shipments per year. Specimens for program C's enzyme program are provided in four 1.0 mL reconstituted lyophilized human serum specimens supplemented with selected analytes. Specimens from program D are provided in three liquid human serum specimens, 24 times per year. Program E has 8 linearly related levels and each level is run twice in a cycle, with 3 cycles a year. Each run consists of 2 samples and has a frequency of every 2 weeks. Therefore each of the 8 levels are analyzed 6 times a year with a total of 48 data points per year.

One report from each EQA/PT program was analyzed. Only major diagnostic instruments were analyzed, and only groups that had more than 10 participant labs.

Data from the general chemistry programs were analyzed, which comprise between 20 and 30 analytes depending on the EQA/PT program: Albumin, alkaline phosphatase, alanine transferase, amylase, AST, direct bilirubin, total bilirubin, calcium, chloride, cholesterol, CO_2, creatinine kinase, creatinine, folate, GGT, glucose, HDL, iron, lactate, LDH, LDL, lipase, magnesium, phosphate, potassium, transferrin, total protein, sodium, triglycerides, uric acid, and urea. Not every analyte was available for analysis. Each EQA/PT program creates its own biochemistry survey with a variety of analytes (another lack of consistency), and not all surveys were available to the author.

Furthermore, the instrument groups listed in each program are different. The smaller EQA/PT programs only list broad manufacturer categories, for example the program C only reports on the 6 major diagnostic manufacturers, not the instrument categories. Often smaller EQA/PT programs can only form groups of sufficient size through broad categorization, while larger EQA/PT programs, with many more participants, can form groups of sufficient size for individual instrument models. For this study, we only focused on the major diagnostic manufacturers and instrument groups with more than 10 participant laboratories (Table 1). As with peer group size, it is commonly expected that groups with less than 10 participants are not of sufficient size to provide reliable results.

Imprecision was determined from the instrument group SD, and CVs were calculated at the level of the instrument group mean. Bias was determined against the all-method consensus mean, when available, or against the peer group method mean, when that was more appropriate. However, as we will see, ultimately the calculation of bias became irrelevant to the study.

CAP/CLIA performance specifications [25] were used as the basis for analytical performance specifications. When a CLIA goal was not available for a particular analyte, for example GGT, Direct Bilirubin and