New criteria for proficiency assessment of clinical laboratories by interlaboratory comparisons

Diogo José da Silva Jerônimo¹² and Elisabeth Costa Monteiro¹

¹Postgraduate Programme in Metrology, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, C.P. 38071, 22451-900, Brazil.
²Control Lab - Controle de Qualidade para Laboratórios, Rio de Janeiro, C.P. 20911-442, Brazil.

beth@puc-rio.br

Abstract. The participation of clinical laboratories in proficiency testing programs is an essential tool for quality management. Using a database provided by an accredited proficiency testing provider, this paper compares the performance of criteria in the assessment of proficiency of participant laboratories for measuring glucose. The analysis considered not only commonly used approaches but also newly proposed performance scores that are independent of the consensus effect, based solely on parameters associated with the assigned value defined by the reference material. The observed discrepancies in the performance rating of laboratories, according to the evaluation criteria employed, allowed the evidencing of inadequacies in the classifications based on scores using parameters affected by the consensus among participants. The present work allowed a better understanding about the impact of criteria applied in proficiency testing programs on the adequacy of evaluation of laboratories' performance, thus contributing to ensure the reliability of clinical diagnostics and therapeutic procedures.

1. Introduction

Along the twentieth century, the increasing role of the clinical laboratory in medical decision making has led to a growing demand for a careful assessment of the reliability of laboratory results [1]. Thus, the accuracy and comparability of measurement results provided by laboratories are essential to guarantee the reliability of diagnoses and treatments [2-4]. Laboratories use a wide variety of techniques to quantify a large number of parameters, so the results may not be comparable across labs [3]. Independent quality assessment for accreditation of medical laboratories, guided by particular requirements for competence and quality described in the international standard ISO 15189 [5], was developed to monitor the performance of individual laboratories. It included the need of participation in appropriate proficiency testing (PT) schemes, aiming at evaluating the measurement capability of participants by interlaboratory comparisons (ILC) operated according to ISO/IEC 17043 [6].

Among the methods described in ISO 13528 to obtain the assigned value of a proficiency test item [7], against which the participants’ data are compared, the most commonly used is the consensus value derived from participants’ results. The appropriateness of this method, however, is highly dependent upon the quality of the results provided by the participants [8]. The effect of outliers, subpopulations, and the substantial increase in the uncertainty of the assigned value determined by the participants' consensus, mainly when few laboratory results are available [9], can, potentially, lead to
misclassification of laboratories’ performance. This risk indicates the importance of using independent criteria, which allow the harmonization of results.

Ideally, PT schemes should provide an estimate of the measurement uncertainty and a statement of the metrological traceability of the assigned value [6]. However, this approach's costs hinder its use, and the procedure conventionally employed is the use of participants’ PT results to calculate the target value and the standard deviation for the proficiency assessment.

The present study compares the results for the PT performance of clinical laboratories for measuring three distinct concentrations of glucose in human serum, evaluated according to six different scoring approaches. Four of the studied performance indicators consist of the most often employed criteria, which incorporate elements based on the consensus value obtained from the outcome of participants. Other two statistical designs are proposed for evaluation, composed of parameters entirely independent of the participants’ measurement results. The six analyzed performance evaluation approaches do not represent costs’ differences for PT implementation.

2. Materials and Methods

The present analysis was performed using the measurement results reported by participating laboratories of a proficiency testing round for glucose determination in samples of lyophilized human serum with three different glucose concentrations (91.08 mg/dL, 301.96 mg/dL, 143.33 mg/dL). The data were provided from a PT round performed by an accredited PT provider. A wide range of analytical systems, consisting of a great variety of reagent/analyzer combinations, were employed by 1737 participating laboratories.

The statistical designs used for performance evaluation of the PT participants were organized into two groups, considering the influence or not of the reported results of laboratories. Group 1 consisted of four combinations of parameters based on recommended criteria described in [7], all of them with some influence from the consensus value (equations 1 to 4).

\[
\begin{align*}
    z\text{-score} &= \frac{x - X_{\text{cons}}}{\sigma_{\text{cons}}} \\
    z_{\text{FL}}\text{-score} &= \frac{x - X_{\text{cons}}}{\sigma_{\text{consFL}}} \\
    z'\text{-score} &= \frac{x - X_{\text{ref}}}{(\sigma_{\text{cons}}^2 + u_{X_{\text{ref}}}^2)^{1/2}} \\
    z'_{\text{FL}}\text{-score} &= \frac{x - X_{\text{ref}}}{(\sigma_{\text{consFL}}^2 + u_{X_{\text{ref}}}^2)^{1/2}}
\end{align*}
\]

Where \( x \) is the result reported by a participant, \( X_{\text{cons}} \), the assigned value obtained from the interlaboratory consensus amongst participants; \( X_{\text{ref}} \), the assigned value derived from the reference value of the provided PT item; \( \sigma_{\text{cons}} \), standard deviation for proficiency assessment derived from the results reported by the participants in the same round; \( \sigma_{\text{consFL}} \), the standard deviation for proficiency, determined using a fixed limit (FL) of 13 % of the assigned value derived from the consensus \( (\sigma_{\text{consFL}} = X_{\text{cons}} \times (FL/2)) \); and \( u_{X_{\text{ref}}} \), the standard uncertainty of the reference value of the provided PT item.

In four of the scoring methods used in this study, the standard deviation for proficiency assessment was established on a fit for purpose value. This approach is strongly recommended in the Harmonized Protocol for the PT in Analytical Chemistry published by the IUPAC (International Union of Pure and Applied Chemistry), and the advantages are widely discussed in the literature [7, 9, 10].

Group 2 consisted of two proposed combinations of parameters entirely independent of the participants’ measurement capability (equations 5 and 6).
\[ z_{FL} = \frac{x - X_{ref}}{\sigma_{refFL}} \]

\[ z'_{FL} = \frac{x - X_{ref}}{\sqrt{(\sigma^2_{refFL} + u_x^2)}} \]

Where \( \sigma_{refFL} \) is the standard deviation for proficiency, determined using a fixed limit (FL) of 13% of the assigned value defined by reference value of the PT item \( \sigma_{refFL} = X_{ref} \cdot (FL/2) \).

After calculation by each one of the six protocols considered, the participants’ results were classified according to their score value outcome, being recognized as acceptable those with \( |z| \leq 2 \), regardless of the statistical method applied [6, 7].

The reference values \( (X_{ref}) \) for each round item, and their standard uncertainties \( (u_{X_{ref}}) \), were obtained from the data provided by the assessment of homogeneity of the reference material, corresponding to each of the three items of glucose concentrations in lyophilized human serum [7, 9].

The assigned values derived from the consensus of participants’ results in a proficiency test round were calculated using the procedures described in the Algorithm A presented in the Annex C.3 of ISO 13528 [7]. The one sample t-test and chi-square test for variance were used to identify differences of acceptance ranges and assigned values between the six performance evaluation procedures.

3. Results

T-test indicated differences between \( X_{cons} \) and \( X_{ref} \) for PT items 1 and 3; and Chi-Square Test showed discrepancies between \( \sigma_{cons} \) and \( \sigma_{consFL} \) for PT items 1 and 2, at a significance level of 0.05.

Figure 1 presents acceptance ranges of reported values for concentration of glucose in human serum associated with item 1 of the PT round, defined according to the six score of the two approaches’ groups (equations 1 to 6). Item 1 values for \( X_{cons} \) (94.05 mg/dL) and \( X_{ref} \) (91.08 mg/dL) are indicated as the assigned values according to the correspondent procedure. The laboratory results represented in Figure 1 include those reported by six ordinary (A to F) and three expert (✦) laboratories. Although laboratories A, B and C were within the acceptable range for the scores whose assigned values were based on the reference value of item 1 \( (X_{ref}) \), the participants have been judged to have an unacceptable performance by the two scores with assigned value derived from the outcome of the PT participants \( (Z-score \ and \ Z-score_{FL}) \), which are false rejections. Likewise, inconsistencies were observed for the performance evaluation results of laboratories D, E and F between the scores based on \( X_{ref} \) to which their reported results were out of acceptance range, and the two approaches based on \( X_{cons} \) that considered measurement result reported as satisfactory. Since for PT item 2 an agreement between \( X_{ref} \) and \( X_{cons} \) was observed, the evaluation discrepancies resulted from the differences between \( \sigma_{cons} \) and \( \sigma_{consFL} \), occurring for scores considering the recommended fixed limits when compared with those that do not apply FL values in the standard deviation for proficiency assessment.

4. Conclusions

This paper presented a comparative performance analysis of distinct scoring criteria employed for evaluation of clinical laboratories by Proficiency Testing schemes, for measuring three concentration items of glucose in human serum. Six different scoring approaches were employed, being two of these newly proposed statistical designs based on parameters entirely independent of the participants’ measurement results. Discrepancies in evaluation rates were observed for PT items for which differences between the assigned value derived from the reference material used in the PT round \( (X_{ref}) \) and obtained by the consensus value from participants \( (X_{cons}) \). If assigned values were coincident, inconsistencies could be produced by differences in the standard deviation for proficiency assessment between scores using or not a recommended fixed limit among their equation parameters.

Thus, acceptable ranges are affected not only by differences between the reference value of the provided PT item and consensus value obtained from the outcome of participants, but also by
discrepancies between values of the assessment standard deviation comprising parameters based on a recommended fixed limit and those derived mainly from the participants’ consensus. Significant differences between parameters used in each performance score approach studied resulted in discrepancies between the acceptable ranges associated with each scoring method, which represents a potential risk of an inappropriate decision of clinical laboratories’ acceptance. Considering the consensus may be biased due to the use of faulty methodology by participants, score inadequacy can be prevented by seeking to apply reliable references to set the assigned value, ideally traceable to SI, utterly independent of the results reported by the PT participating laboratories.

**Figure 1.** Acceptance ranges for reported results to item 1 of glucose concentration, according to the six scoring methods associated with the equations (1) to (6). Results reported by six ordinary laboratories (A to F) and three experts (♦) are shown on each adequacy range. Evaluation discrepancies are indicated as false approval (+) and false rejection (×).

**Acknowledgement**
The authors acknowledge the Control Lab management for providing the data analyzed in this study.

**References**
[1] Plebani M, Sciacovelli L and Aita A 2017 *Clin Lab Med* **37** 187–205
[2] Costa Monteiro E 2007 *Metrologia e Instrumentação*, 6 6-12
[3] Delatour V, Lalere B, Saint-Albin K, Peignaux M, Hattchouel J M, Dumont G, De Graeve J, Vaslin-Reimann S and Gillery P 2012 *Clin Chim Acta* **414** 182–184
[4] Costa Monteiro E and Leon L F 2015 *Journal of Phys. Conf. Series* **588** 012032
[5] ISO 15189:2012, Medical laboratories – Particular requirements for quality and competence
[6] ISO/IEC 17043:2010, Conformity assessment — General requirements for proficiency testing
[7] ISO 13528:2015, Statistical methods for use in proficiency testing by interlaboratory comparisons
[8] De Bièvre P 2016 *Accred Qual Assur* **21** 167–169
[9] Thompson M, Ellison S L R and Wood R 2006 *Pure Appl. Chem* **78**(1) 145–196
[10] Ehrmeyer S S and Laessig R H 1987 *Clinical Chemistry* **33**(10) 1901–1902