Harmonization and transferability of performance assessment: experience from four serum aluminum proficiency testing schemes

Josiane Arnaud, Isabelle Coté, Patrick J. Parsons, Marina Patriarca, Andrew Taylor & Cas Weykamp

Accreditation and Quality Assurance
Journal for Quality, Comparability and Reliability in Chemical Measurement

ISSN 0949-1775

Accred Qual Assur
DOI 10.1007/s00769-014-1049-4

Springer

Accred Qual Assur
DOI 10.1007/s00769-014-1049-4

PRACTITIONER'S REPORT

Harmonization and transferability of performance assessment: experience from four serum aluminum proficiency testing schemes

Josiane Arnaud · Isabelle Coté · Patrick J. Parsons · Marina Patriarca · Andrew Taylor · Cas Weykamp

Received: 18 December 2013 / Accepted: 3 April 2014

Abstract Proficiency testing schemes monitor laboratory performance and provide a stimulus for improvement in accuracy. Where several schemes operate in the same analytical sector, there are risks that assessments of performance may be in conflict. Performance assessment for the determination of trace elements such as aluminum in serum is particularly important due to the high risk of contamination and therefore erroneous results. The objectives of this work were (1) to compare several mathematical models to establish a predefined standard deviation for proficiency assessment and (2) to evaluate the influence of instrumental methods and proficiency testing scheme on the assessment of performance for serum aluminum measurements. For this purpose, three samples were sent to the participants of four proficiency testing schemes. Assigned values were calculated according to algorithm A according to ISO 13528 and standard deviation for proficiency assessment according to three methods based on individual variability, state of the art or previous proficiency testing results. The method based on individual variability produced a more stringent standard deviation compared to analytical imprecision based on the state of the art. The instrumental methods gave similar results, whereas significant differences were observed between the four proficiency testing schemes indicating that harmonization of the standard deviation for proficiency assessment fails to allow transferability from one proficiency testing scheme to another and that additional factor(s) contribute to variability in performance assessment.

Keywords Proficiency testing · Serum aluminum · Evaluation of performance

All the authors are the members of the thematic NETWORK "Organisers of external quality assessment/proficiency testing schemes related to occupational and environmental medicine."

J. Arnaud (✉)
Département de Biochimie, Toxicologie et Pharmacologie, CHU de Grenoble, Grenoble, France
e-mail: JArnaud@chu-grenoble.fr

I. Coté
Centre de Toxicologie, Institut National de Santé Publique du Québec, Quebec, Canada

P. J. Parsons
Laboratory of Inorganic and Nuclear Chemistry, New York State Department of Health, Wadsworth Center, Albany, NY, USA

M. Patriarca
Department of Public Veterinary Health and Food Safety, Istituto Superiore di Sanità, Rome, Italy

A. Taylor
Department of Clinical Biochemistry, Royal Surrey County Hospital, Guildford, UK

C. Weykamp
MCA Laboratory, Queen Beatrix Hospital, Winterswijk, The Netherlands

Introduction

Trace element proficiency testing (PT) schemes monitor laboratory performance and provide a stimulus for improving accuracy. However, assessing a participant's performance depends on the statistical method in use and the criteria used by the PT provider to determine the assigned values and the standard deviation for proficiency assessment ($\hat{\sigma}$). As a consequence, the performance of the same laboratory could be considered acceptable in one

Published online: 24 April 2014

scheme but questionable or unacceptable in another [1–4]. It may be a mandatory requirement for laboratory accreditation [4–6] to participate successfully in a PT program. However, harmonized accreditation and mutual comparability cannot be achieved when different performance standards are set by PT providers. Previously, the CoEPT study evaluated the comparability of data between PTs in the field of water, food, soil and occupational hygiene [4]. The NETWORK of organizers of external quality assessment schemes in Occupational and Environmental Laboratory Medicine members collaborate to define common acceptable levels of performance in the field of clinical biochemistry and occupational health, taking into account both clinical and analytical issues. The NETWORK has published recommended quality specifications for aluminum (Al) in serum [7] and shown how to use these to derive $\hat{\sigma}$ [2]. The objectives of this study conducted by the NETWORK were (1) to compare several mathematical models to establish a predefined $\hat{\sigma}$ and (2) to evaluate the influence of instrumental methods and PT schemes on the assessment of performance for concentration measurements of aluminum in human serum.

Materials and methods

Samples and participants

A pool of human serum with a low endogenous concentration of Al, Cu, Se and Zn was used for this study; two additional serum pools were prepared by spiking the first with known amounts of Al, Cu, Se and Zn to produce elevated concentrations. For Al, the spikes were 1.72 μmol L^{-1} and 3.43 μmol L^{-1} (46.5 μg L^{-1} and 93.0 μg L^{-1}). These three pools provided common samples that were sent at ambient temperature to all participants of the French (FR, number of participants $n = 21$), Italian (IT, $n = 25$), New York State (NY, $n = 30$) and United Kingdom (UK, $n = 61$) PT schemes for serum trace elements. These PTs were not focused on Al alone. The FR, IT and NY PTs provided samples for Al, Cu, Se and Zn determinations whereas UK provided samples for Al, Au, Cu, Se and Zn concentration measurements. The participants were instructed to analyze the samples and submit their results within 4–6 weeks according to the usual scheme policy. In addition, participants reported instrumental details for the following methods: electrothermal atomic absorption spectrometry with Zeeman background correction (ETAAS-Z); or with unspecified background correction (ETAAS); inductively coupled plasma atomic/optical emission spectrometry (ICP-OES); inductively coupled plasma mass spectrometry (ICP-MS). Table 1 lists the instrumental techniques applied and the number of participants using them in the different schemes.

Table 1 Number of participants of the four different PT schemes using a specified instrumental method; amounting to a total of 137 participants

Instrumental method	FR	IT	NY	UK
ETAAS	3	5	2	19
ETAAS-Z	12	15	11	27
ICP-OES	2	3	2	2
ICP-MS	2	2	14	8
Other or not documented	2	0	1	5

Statistics

Individual results from the different schemes were pooled for statistical analysis. Consensus robust values for mean (x^*) and standard deviation (s^*) were calculated according to algorithm A [8] from the results reported to each scheme as well as for all the participants ($x^*_{overall}$, $s^*_{overall}$). The standard uncertainty (u) was calculated according to Eq. 1 [8]:

$$u = (1.23s^*)/\sqrt{n} \quad (1)$$

Participant Z-scores (Z) were determined using Eq. 2

$$Z = (x - x*_{overall})/\hat{\sigma}_{overall} \quad (2)$$

where x denotes the individual value and $\hat{\sigma}$ is based on the following three different approaches:

(1) Minimal performance based on individual variability as proposed by Fraser [9] and using the data of Nordal et al. [10]. $\hat{\sigma}$ previously calculated by the NETWORK for Al concentration in serum [7] was 0.09 μmol L^{-1} (2.5 μg L^{-1}) or 10 % whichever was greater.

(2) Horwitz et al. [11] criterion, based on the relationship between concentration and $\hat{\sigma}$

$$s_R = 0.02\ C^{0.85} \quad (3)$$

where C is the mass fraction in g g^{-1} and s_R the standard deviation for performance assessment in g g^{-1}. Al relative atomic mass of 26.98 and serum density of 1.024 g mL^{-1} were used to convert the unit.

(3) Thompson's approach [12, 13], based on data obtained by the different PT schemes.

$$s_R = \sqrt{\alpha^2 + (\beta C)^2} \quad (4)$$

where α is the constant standard deviation at concentrations close to the detection limit, and β is the constant relative standard deviation (CV of reproducibility) at high concentrations. The data obtained for the two previous years ($n = 85$ serum samples from 0.45 μmol L^{-1} to 10.4 μmol L^{-1}) were

used for the calculation of α and β. The relationships between $\log_{10} CV$ or $\log_{10} s_R$ and $\log_{10} C$ are presented in Fig. 1. Obtained α and β values were 2.09 ng g^{-1} and 0.156, respectively.

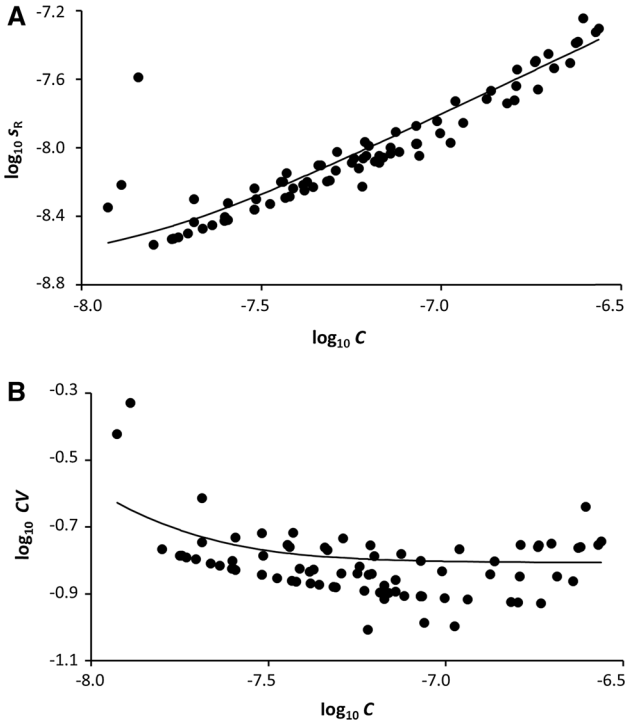

Fig. 1 Robust standard deviation (A) and reproducibility *CV* (B) versus robust mean of Al mass fractions in serum obtained by the proficiency testing schemes in the two previous years. The solid line represents the "uncertainty function" (Eq. 4) fitting the experimental data points [12]

Table 3 Standard deviation for proficiency assessment ($\hat{\sigma}$) according to different models, in µmol L^{-1}

	Sample 1	Sample 2	Sample 3
$\hat{\sigma}$ Fraser [9], Taylor et al. [7]	0.09	0.23	0.40
$\hat{\sigma}$ Horwitz et al. [11], Eq. 3	0.16	0.56	0.90
$\hat{\sigma}$ Thompson [12], Eq. 4	0.12	0.37	0.63

A chi-square test was used for the comparison of qualitative data, whereas Kruskal–Wallis or Mann–Whitney tests were performed for continuous variables. $p < 0.05$ was considered as significant.

Results

Comparison of the three methods to establish the standard deviation for proficiency assessment ($\hat{\sigma}$)

The x^*_{overall} (0.55 µmol L^{-1}, 2.29 µmol L^{-1} and 4.02 µmol L^{-1}) were taken as the assigned values (Table 2). Then, the $\hat{\sigma}$ values were calculated according to the three different approaches (Table 3). The most stringent criteria corresponded to the approach based on individual variability [7, 9], whereas those which took into account only the concentration [11] were less stringent. Observed s^* were larger than the $\hat{\sigma}$ values calculated according to the recommendations of the NETWORK [7] and smaller than those calculated according to the Horwitz et al. [11] method.

Participant's performances evaluated by *Z*-score are indicated in Table 4. Numerous unsatisfactory results

Table 2 Robust means (x^*), robust standard deviations (s^*) and standard uncertainties (u) for serum Al concentrations in the three samples according to instrumental methods and the four PT schemes

Sub-groups	Sample 1				Sample 2				Sample 3			
	n	x^*	s^*	u	n	x^*	s^*	u	n	x^*	s^*	u
overall	137	0.55	0.14	0.015	135	2.29	0.33	0.035	135	4.02	0.58	0.062
ETAAS	29	0.56	0.20	0.045	29	2.36	0.36	0.084	28	4.10	0.65	0.153
ETAAS-Z	65	0.54	0.15	0.023	64	2.29	0.35	0.054	65	3.98	0.72	0.112
ICP-OES	9	0.52	0.15	0.062	9	2.20	0.44	0.182	8	3.77	0.59	0.261
ICP-MS	26	0.54	0.14	0.034	25	2.24	0.22	0.056	26	4.05	0.25	0.061
FR	21	0.47	0.10	0.028	21	2.19	0.39	0.107	20	3.86	0.51	0.142
IT	25	0.55	0.22	0.055	25	2.32	0.43	0.107	24	3.87	1.08	0.276
NY	30	0.54	0.16	0.037	30	2.26	0.31	0.071	30	3.96	0.44	0.100
UK	61	0.58	0.14	0.022	59	2.32	0.25	0.041	61	4.15	0.51	0.082

All values in µmol L^{-1} (except n)

n number of participants

x^* and s^* calculated according to algorithm A [8]

u calculated using Eq. 1

Table 4 Percentages of questionable and unsatisfactory *Z*-scores[a] ($2 < |Z| \leq 3$ and $|Z| > 3$, respectively) when applying the three evaluation models; the scores obtained with all samples by all participants are included and itemized by instrumental method and the four PT schemes

Sub-groups	Fraser [9], Taylor et al. [7]		Horwitz et al. [11]		Thompson [12]	
	Questionable (%)	Unsatisfactory (%)	Questionable (%)	Unsatisfactory (%)	Questionable (%)	Unsatisfactory (%)
Overall	11.1	13.8	3.4	4.4	6.1	5.2
ETAAS	8.1	19.8	5.8	8.1	8.1	11.6
ETAAS-Z	13.4	14.9	3.6	3.1	10.8	5.7
ICP-OES	11.5	15.4	0.0	11.5	3.8	11.5
ICPMS	7.8	3.9	2.6	1.3	3.9	3.9
FR	11.3	12.9	8.1	0.0	8.1	6.5
IT	16.2	29.7	2.7	10.8	16.2	13.5
NY	8.9	7.8	1.1	2.2	10.0	2.2
UK	9.9	10.5	3.3	4.4	4.4	6.6

[a] Values of $x^*_{overall}$ and $\hat{\sigma}_{overall}$ were used for the calculation of *Z*-scores according to Eq. 2

($|Z| > 3$) were observed with most showing a large positive bias as evidenced by the percentage of *Z*-scores higher than 3 (8.1 % with Fraser's approach, 3.9 % with Horwitz et al. approach and 5.2 % with Thompson's approach).

Influence of instrumental method on performance assessment

The distribution of instrumental methods used by participants differed significantly (Table 1, $p < 0.0001$) between PT schemes. The relevance of the larger proportion of laboratories using ICP-MS in the NY scheme is discussed below.

The x^* and s^* values did not differ significantly according to the instrumental method (Table 2; $p > 0.53$ whatever the sample). However, assigned values of samples 2 and 3 for ICP-OES users were lower than the range $x^*_{overall} \pm 2u$ (Fig. 2, Table 2).

The performance classified as satisfactory ($-2 \leq Z \leq 2$), questionable ($2 < |Z| \leq 3$) or unsatisfactory ($|Z| > 3$) did not vary according to the instrumental method used ($p \geq 0.05$ whatever the approach used [7, 9, 11, 12]). This lack of influence by method was confirmed by the comparison of *Z*-scores according to instrumentation ($p \geq 0.38$ whatever the approach used).

Influence of proficiency testing on performance assessment

Regarding the four PT schemes, a significant difference in the concentration x^* was observed with sample 1 (Table 2, $p = 0.04$) but not with samples 2 ($p = 0.45$) and 3 ($p = 0.16$). However, the three x^* values obtained by the FR PT scheme (sample 1: 0.47 µmol L^{-1}; sample 2: 2.19 µmol L^{-1} and sample 3: 3.86 µmol L^{-1}) were unexpectedly, significantly and systematically lower than the $x^*_{overall}$ minus their expanded uncertainty (U, $k = 2$) (Fig. 2, Table 2).

The performances by scheme were significantly different (Table 4). Expressed as satisfactory, questionable and unsatisfactory classes, p varied from 0.01 (Horwitz et al. [11] approach, to 0.0003 using the approach based on individual variability [7], and 0.041 when using the $\hat{\sigma}$ calculated according to Thompson [12, 13]. These results were confirmed by the comparison of the *Z*-scores ($p = 0.002$ whatever the approach used).

Discussion

Comparison of three methods to establish standard deviation for proficiency assessment

Fraser's approach, based on intra- and inter-individual variability, reflects fitness for the purpose of the assay and should be preferred as far as possible. However, calculation of $\hat{\sigma}$ is based on data dealing with individual variability and to our knowledge only one paper has been published on this subject for aluminum concentrations in serum [10]. This study was used by the NETWORK to define the $\hat{\sigma}$ for Al [7]. For other trace elements, data for intra- and inter-individual variability are inconsistent [14, 15] as many factors may influence the intra-individual (e.g., frequency and time of sampling, food intake, medication) and inter-individual variability (e.g., age, gender, race, education, occupation, life style). Therefore, while the $\hat{\sigma}$ may depend on the design and results of the published studies, recent reliable data appear to be reasonably consistent [16]. On the contrary, the two other approaches are based solely on analytical performance and require that method performance should be fit for the purpose of clinical investigations. The $\hat{\sigma}$ values based on individual variability

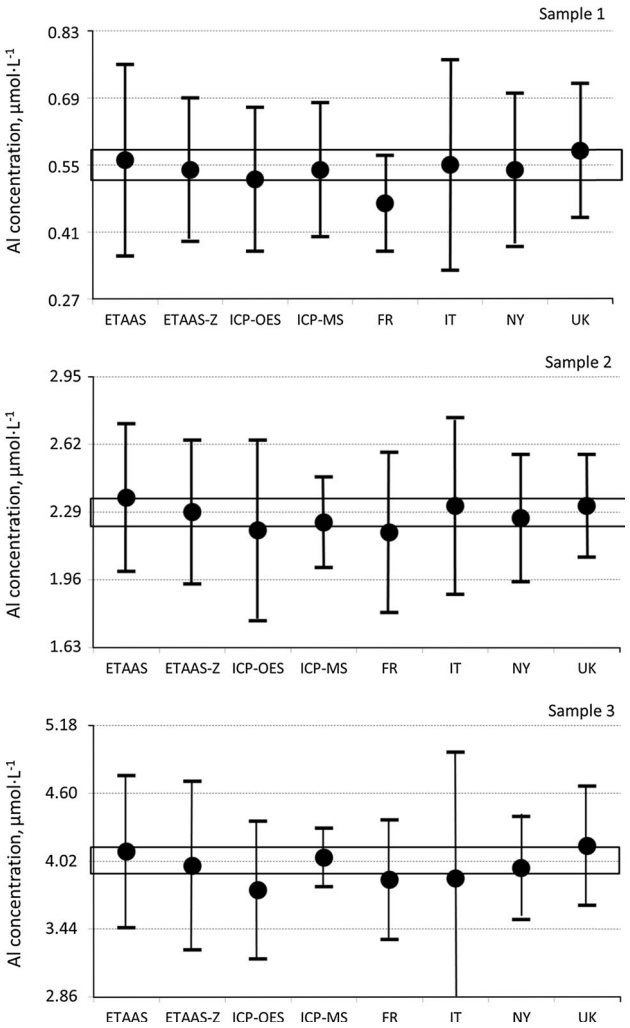

Fig. 2 Robust means (x^*), robust standard deviation (s^*) and expanded uncertainties (U, $k = 2$) obtained for Al concentrations in the three serum samples according to instrumental methods and the four PT schemes. The figures on Y scale correspond to $x^*_{overall} \pm ms^*$ with $m = 1, 2$. Error bars indicate $x^* \pm 1\ s^*$ individually for each group. *Rectangle* indicates the range $x^*_{overall} \pm U$ ($k = 2$)

[7], might appear to be too stringent as fewer than 80 % of participants obtained satisfactory results. However, this initial conclusion must be reconsidered as 88.3 % of ICP-MS users and 83.3 % of the NY scheme participants obtained satisfactory results (see below). The $\hat{\sigma}$ values given according to the Horwitz et al. [11] formula, based on concentration, are too large as they are greater than the consensus s^* and therefore are not fit for purpose as they do not provide any stimulus for overall analytical improvement. The approach proposed by Thompson [12], based on recent analytical performance provided $\hat{\sigma}$ values that were similar to the observed s^*. This is not surprising as the calculations were based on data from the same laboratories over a previous period of time. This approach remains the best one when no information is available on individual variability. The risk in using analytical approaches is failure to identify laboratories that may report misleading results, to the detriment of patients. This is particularly important for the measurement of serum aluminum concentrations as our data showed that unsatisfactory results were generally positively biased suggesting that contamination was not under adequate control in poorly performing laboratories.

Influence of instrumental method on performance assessment

The instrumental methods used by the participants differed to some extent according to specific schemes with a greater proportion of ICP-MS users in the NY PT program. The most frequently used method was ETAAS-Z. No statistically significant differences according to instrumental methods were observed which suggests that results from the four most common techniques are equivalent. While it is noteworthy that as a group, ICP-MS had better precision than the other instrumental methods, some caution is warranted in "over-interpreting" this as ETAAS is more likely to be implemented differently by participant laboratories. In addition, the number of ICP-OES users was relatively small, so the observed low bias must be confirmed.

Influence of proficiency testing scheme on performance assessment

Although statistically significant differences were noted in the x^* values by schemes only for sample 1, the differences in participants' results are sufficient to modify the assigned value and therefore the range of satisfactory results. Indeed, taking sample 3 and $\hat{\sigma}$ based on individual variability as an example, a result of 4.90 µmol L^{-1} would be considered as satisfactory by the UK schemes while being flagged as questionable by the three other PTs. Therefore, while harmonization of $\hat{\sigma}$ is an improvement in performance assessment it is not sufficient to completely assure transferability of results from one scheme to another and consideration should also be given to determination of the assigned value. In addition, differences in s^* suggest that the approach of Thompson [12] is less transferable. Using this approach, the $\hat{\sigma}$ is based on previous PT data. Consequently, the greater s^* observed in the IT program translates into a larger, i.e., less challenging $\hat{\sigma}$.

As indicated in Table 4, performance varies according to PT schemes with the best observed in the NY PT program and the worst in the IT PT scheme. A possible reason for the better performance in the NY PT may be that these participants are subject to mandatory clinical laboratory standards specifically for trace element analysis, including

procedures for contamination control, and laboratories are routinely inspected every 2 years [17]. In addition, failure to achieve successful performance in the NY PT program results in the participant being required to cease all patient testing, conduct a root cause analysis and submit an acceptable corrective action plan to the accrediting authorities [18]. A similar observation was noted for UK-based participants within the UK scheme. These laboratories are also subject to independent scrutiny and poor performers are reported to health authorities. Performance for this group of participants was shown to be superior to that of the non-UK participants in the same scheme [19].

Conclusions

Despite the paucity of studies dealing with individual variability of Al concentrations in serum, the $\hat{\sigma}$ proposed by the NETWORK in 2002 based on individual variability [7], although quite stringent, remains the preferred approach for harmonizing the performance evaluation by different PT schemes. The calculation of $\hat{\sigma}$ by this approach does not depend on PT participant performances in contrast to Thompson's approach [12]. In addition, this approach is fit for the purpose of clinical investigations. Nonetheless, a standardized protocol for studying individual variability is greatly needed. Our results demonstrate that harmonization of $\hat{\sigma}$ is insufficient to ensure transferability of performance evaluation from one scheme to another if the assigned value is determined as the x^* as this depends on the participant population i.e., the number of participants, instrumental methods employed and the effectiveness of contamination control procedures. Similar observations were reported in the CoEPT for other matrices (air, water, soil and food) [4]. To achieve transferability, it would also be necessary for scheme providers to show that assigned values have metrological traceability [20] and to educate participants.

References

1. Christensen JM, Olsen E (1996) Comparison of some European external quality assessment schemes in the field of occupational and environmental medicine. Ann Ist Super Sanita 32:285–294
2. Taylor A (2011) Quality assessment of measurement. J Trace Elem Med Biol 25S:S17–S21
3. Mann I, Brookman B (eds) (2011). Selection, use and interpretation of proficiency testing (PT) schemes. 2nd edn. Eurachem
4. CoEPT project (2002) Eptis. http://www.eptis.bam.de/en/further_pt_info/coept_project. 20 Aug 2013
5. ISO 15189 (2012) Medical laboratories: requirements for quality and competence, 3rd edn. ISO/TC212, ISO, Geneva, Switzerland
6. ISO/IEC 17025 (2005/cor1: 2006) General requirements for the competence of testing and calibration laboratories, 2nd edn. ISO/CASCO, ISO, Geneva, Switzerland
7. Taylor A, Angerer J, Claeys F, Kristiansen J, Mazarrasa O, Menditto A, Patriarca M, Pineau A, Schoeters I, Sykes C, Valkonen S, Weykamp C (2002) Comparison of procedures for evaluating laboratory performance in external quality assessment schemes for lead in blood and aluminum in serum demonstrates the need for common quality specifications. Clin Chem 48:2000–2007
8. ISO 13528 (2005) Statistical methods for use in proficiency testing by interlaboratory comparison. ISO/TC69/SC 6, ISO, Geneva, Switzerland
9. Fraser CG (1999) General strategies to set quality specifications for reliability performance characteristics. Scand J Clin Lab Invest 59:487–490
10. Nordal KP, Dahl E, Thomassen Y, Brodwall EK, Halse J (1988) Seasonal variations in serum aluminum concentrations. Pharmacol Toxicol 62:80–83
11. Horwitz W, Kamps LR, Boyer KW (1980) Quality assurance in the analysis of foods and trace constituents. J Assoc Off Anal Chem 63:1344–1354
12. Thompson M (1988) Variation of precision with concentration in an analytical system. Analyst 113:1579–1587
13. Cote I, Robouch P, Robouch B, Bisson D, Gamache P, LeBlanc A, Dumas P, Pedneault M (2012) Determination of the standard deviation for proficiency assessment from past participant's performances. Accred Qual Assur 17:389–393
14. Gallagher SK, Johnson LK, Milne DB (1989) Short-term and long-term variability of indices related to nutritional status. I: Ca, Cu, Fe, Mg, and Zn. Clin Chem 35:369–373
15. Gonzales-Revalderia J, Garcia-Bermejo S, Menchen-Herreros A, Fernandez-Rodriguez E (1990) Biological variation of Zn, Cu, and Mg in serum of healthy subjects. Clin Chem 36:2140–2141
16. Arnaud J, Weber J-P, Weykamp C, Parsons PJ, Angerer J, Mairiaux E, Mazarrasa O, Valkonen S, Menditto A, Patriarca M, Taylor A (2008) Quality specifications for the determination of copper, zinc, and selenium in human serum or plasma: evaluation of an approach based on biological and analytical variation. Clin Chem 54:1892–1899
17. Trace elements (2008) http://www.wadsworth.org/labcert/clep/files/BLLE_TREL_EP_effective_07_01_10.pdf
18. CLIA (2003) Standards and certification: laboratory requirements (42 CFR 493). GPO, cdc, Atlanta, GA, USA
19. Taylor A, Walker AW (1992) Measurement of aluminium in clinical samples. Ann Clin Biochem 29:377–389
20. ISO/IEC 17043 (2010) Conformity assessment: general requirements for proficiency testing. ISO/CASCO, ISO, Geneva, Switzerland