Clinical Biochemistry 57 (2018) 12–17

Contents lists available at ScienceDirect

Clinical Biochemistry

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

Performance specifications and six sigma theory: Clinical chemistry and industry compared

Wytze P. Oosterhuis*, Michel J. Severens

Zuyderland Medical Center, Department of Clinical Chemistry and Hematology, Sittard, The Netherlands

ABSTRACT

Analytical performance specifications are crucial in test development and quality control. Although consensus has been reached on the use of biological variation to derive these specifications, no consensus has been reached which model should be preferred. The Six Sigma concept is widely applied in industry for quality specifications of products and can well be compared with Six Sigma models in clinical chemistry. However, the models for measurement specifications differ considerably between both fields: where the sigma metric is used in clinical chemistry, in industry the Number of Distinct Categories is used instead. In this study the models in both fields are compared and discussed.

1. Introduction

How to develop and apply quality control in clinical chemistry is a problem that is intensively debated lately as the method to determine quality specifications is still controversial [1–4]. The conventional total error theory has been questioned. One of the reasons is the general recognition of flaws in the definition of the permissible (or allowable) total analytical error (pTAE) when based on biological variation [2,3].

In short the main flaws concern firstly, that both maxima of permissible bias and imprecision are added to obtain pTAE; these two maximum permissible errors are derived under the mutually exclusive conditions of zero bias and zero imprecision, respectively. Secondly, the maximum permissible bias was derived from a model for diagnosis, while this bias term is applied for monitoring that requires a stricter specification.

Consensus is lacking for a quality control model that integrates concepts of error, measurement uncertainty and Six-Sigma and that offers an unflawed solution for the issue of quality limits based on biological variation [5].

It is well known that the Six-Sigma concept originates from industry, introduced by engineers Smith and Harry while working at Motorola in 1986, and adopted by General Electric in 1995. The Six Sigma concept has also been introduced in the clinical laboratory in 2000 [6] and clinical chemistry [7]. However, some of the methods that are routinely applied in industry are generally unknown in the medical laboratory. It might be interesting to take notice of these procedures outside the field of clinical chemistry. In industry, quality control of manufacturing processes has many concepts and methods in common with quality control in clinical chemistry, and we might learn from the solutions that have shaped the routine procedures there.

1.1. The six-sigma model and quality control in industry

In automobile industry, many components are assembled and the quality requirements are high: when a component has a quality below limits this could for several reasons lead to substantial financial loss. The quality costs could be represented by the number of items outside specification multiplied by the cost of rework or scrap. In exceptional cases, however, this has the potential of severely disrupting the manufacturing process. It has even been argued that any item manufactured away from the target – even within specified limits - would result in loss to the customer [8]. Such losses would inevitably find their way back to the manufacturer and that by working to minimise them, manufacturers would enhance brand reputation, win markets and generate profits. These are all reasons for industry to apply rigorous quality control procedures.

1.2. Basic six-sigma concepts

What are the concepts used in industry, based on the Six-Sigma model? As an example, let us assume we have a product or component with a certain average width of 10 (arbitrary units, suppose this is equal to the target value). The different samples of this product will have some variation around this average value due to small variations (PV, process variation) in the production process. In our example (Fig. 1) the standard deviation (SD) is assumed to be equal to 2. Product

* Corresponding author at: Zuyderland Medical Center, Department of Clinical Chemistry and Hematology, Heerlen, The Netherlands.
E-mail address: w.oosterhuis@zuyderland.nl (W.P. Oosterhuis).

https://doi.org/10.1016/j.clinbiochem.2018.04.002
Received 20 January 2018; Received in revised form 26 March 2018; Accepted 3 April 2018
Available online 11 April 2018

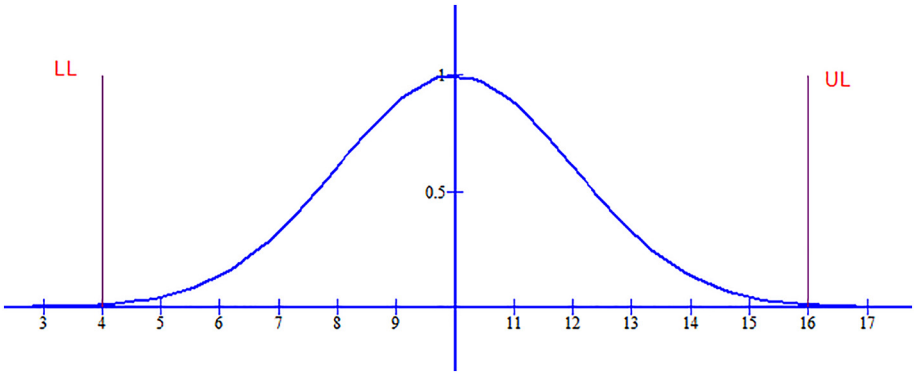

Fig. 1. Six-Sigma model. Schematic representation of an industrial product of 3 sigma quality, with a mean value of 10 arbitrary units and variation (SD) of 2 units and lower and upper performance limits (LL and UL) at ± 3 SD. In this case defects on either side are expected to be 1350 DPMO. A Six-Sigma quality would require a SD (or σ-value) equal to 1.0.

specifications expressed as upper- and lower performance limits (UL, LL) will be determined based on the application of the product (here: LL = 4, UL = 16). The observations outside the limits are considered defects, and are counted as defects per million opportunities (DPMO). If we assume that the limits are at 3.0 SD, and the distribution is Gaussian, 0,135% of the observations will be outside each of the limits. Taking both sides into account, the total DPMO will in this case be equal to 2700 (Fig. 1). (note that this represents the so-called "short-term" sigma in the Six Sigma model, where the 1.5 SD shift is not included as is in the "long-term" sigma). Manufacturers will try to keep product variation as low as possible: in the Six-Sigma concept the goal is a product variation (SD or σ-value) that is 1/6 of the interval between target and performance limit. In the example, product variation of Six-Sigma quality would require a SD (or σ) equal to 1.0.

1.3. Measuring system analysis (MSA) and gage repeatability & reproducibility

The product should fulfil the pre-set quality specifications and a measurement procedure is needed for proper quality control purposes. The measurement should be of sufficient accuracy to distinguish a good product from a bad product. For the specifications of the measurement the procedure of the Automotive Industry Action Group (AIAG) is often used where the methods of measurement system analysis are described in detail [9].

Gage Repeatability & Reproducibility, commonly known as a Gage R&R, is a statistical method used in industrial process control to measure the variation related to a measuring procedure and the subsequent effectiveness of the instrument (in general named *gage*) to be used as a measuring tool. This is part of the Six-Sigma approach as applied in industry.

There are, as the name suggests, two main components that comprise the Gage R&R:

Repeatability, related to the ability of the instrument to give consistent results under identical conditions. It represents the inherent variation of the equipment itself. Repeatability is usually called the "within appraiser" or "within system" variation.

Reproducibility, or the ability of the gage to provide repeated results regardless of the operator performing the test (variation among operators) and is usually referred to as "(between) appraiser variation". This is often applicable when the manual equipment is influenced by the operators' skill. However, it is does not directly apply to automated systems where reproducibility is considered as the average variation between conditions or between systems of measurement. In that case it is called the "between system" variation.

Repeatability and reproducibility are the short-term components of variation

$$\sigma^2_{GageR\,\&\,R} = \sigma^2{}_{\text{repeatability}} + \sigma^2{}_{\text{reproducibility}}$$

Gage R&R is an estimate of the combined variation of repeatability and reproducibility and represents the best (or basic) performance characteristic of the measuring system. Gage R&R studies determine how much of the observed process variation is due to measurement system variation. Gage R&R is a well-known procedure for evaluating measurement systems. A typical study utilizes 3 operators for one measuring device that is measuring a single characteristic, each operator measuring 10 parts in duplicate and covering the whole measuring range [9,10].

In a good measurement system, repeatability and reproducibility should be in proportion with the part-to-part variation, meaning that the measurement system can effectively distinguish differences between the measured characteristics of the parts.

Gage R&R defines the measurement *capability*. Simple capability includes the components of repeatability and reproducibility with uncorrected bias or linearity. An estimate of measurement capability, therefore, is an expression of the expected error for defined conditions including the time window, scope and range of the measurement system.

As with the manufacturing process performance, measurement system *performance* is the net effect of all significant and determinable sources of variation over a longer time period. Performance quantifies the long-term assessment of combined measurement errors (random and systematic). Therefore, performance includes the long-term error components of both capability (short-term errors) and stability (long-term error). The total performance of a measurement system is considered to be composed of the additional component:

$$\sigma^2{}_{\text{performance}} = \sigma^2{}_{\text{GRR}} + \sigma^2{}_{\text{stability}}$$

stability represents the change in bias or drift over time.

Just as short-term *capability*, long-term *performance* is always associated with defined measurement conditions and time period. Measuring system performance is the net effect of all significant and determinable sources of variation over time. Performance quantifies the long-term assessment of combined measurement errors (random and systematic).

Note that in the following model only the short-term Gage R&R is used.

1.4. Quality criteria according to the AIAG

The criteria as to whether a measuring system's variability is satisfactory are dependent upon the ratio of the manufacturing production process variability and the measuring system variation.

Following this procedure, the ratio of the process (or part) variability (PV or σ_{part}) to the short time variation (repeatability and reproducibility, Gage R&R or $\sigma_{\text{measurement}}$) is calculated. See Table 1 for interpretation of these ratio's. Central here is the concept of *"Number of Distinct Categories"* (NDC). The ratio of the process variability and the short time variation (Gage R&R) is used to calculate the Number of Distinct Categories (NDC).

Table 1
Interpretation of the levels of the Number of Distinct Categories (NDC).

NDC	Decision	Comments
> 14	Generally considered to be an acceptable measuring system	Recommended, especially useful when trying to sort or classify parts or when tightened process control is required.
4–14	May be acceptable for some applications	Decision should be based upon, for example, importance of application measurement, cost of measuring device, cost of rework or repair. Should be approved by customer.
< 4	Considered to be unacceptable	Every effort should be made to improve the measuring system. This condition may be addressed by the use of an appropriate measurement strategy; for example, using the average result of several readings of the same part characteristic in order to reduce final measurement variation.

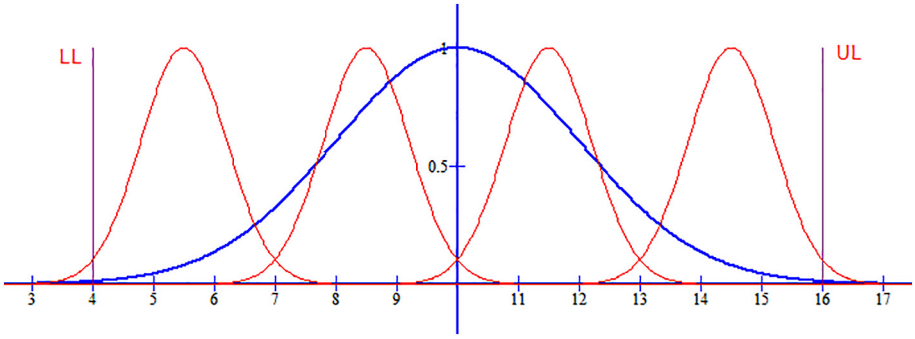

Fig. 2. Example of industrial parts with an average of 10 (arbitrary units) and a part variability (PV) of 2 (central curve). The four curves represent the measurement variability. The Number of Distinct Categories (NDC) of the measuring system is depicted and equals 4 (the limit of acceptability).

$$NDC = \sqrt{2}\frac{\sigma_{part}}{\sigma_{measurement}}$$

The NDC is a measure used in measuring system analysis and refers to the number of distinct (non-overlapping confidence intervals) product categories that can be distinguished by a measuring system.

In our example (Fig. 2), assuming a hypothetical part variation PV = 2, a measurement variation = 0.7, the NDC can be calculated as:

NDC = 1.4 (2/0.7) = 4.0

It important to note that the quality of measurement is generally not expressed as sigma-score (see note 1).

1.5. Performance specifications in clinical chemistry

How does the model as generally applied in industry compare to the methods used in clinical chemistry? The industrial product is best compared in the model with the analyte in the patient sample, as these are both subject to the measurement procedure. The part-to-part variation corresponds with the biological variation within the patient in monitoring applications, or the combined within- and group variation in diagnostic applications. It could, however, be considered to include all sources of pre-analytical variation.

1.5.1. Conventional linear model (total error model)

The performance specification is generally expressed as maximum permissible (allowable) analytical variation (pCV_a) and not directly expressed as a specification limit. According to the commonly accepted performance specifications this is expressed as [11]:

$pCV_a < 0.5CV_I$

Where CV_I represents the within subject biological variation. The variation is commonly expressed as CV.

Example: Creatinine (Fig. 3):

CV_I = 5.95%(8.15 μmol/L)

CV_G = 14.7%

CV_a = 2.55%(at the level of 137 μmol/L)

SD_a = 3.5 μmol/L

These values are based on the biological variation database [12] (*) or local values (#) (Table 2).

In the Six Sigma model, we define upper and lower performance limits. How to deal with this in case we start with a permissible (or allowable) CV_a and how to derive the performance limits? The conventional model is defined as permissible Total Error (pTE):

$pTAE\% = 0.25(5.95^2 + 14.7^2)^{0.5} + 1.65(0.5*5.95) = 8.87\%$

pTAE = 0.0887∗137 = 12.2 μmol/L

UL = 137 + 12.2 = 149.2 μmol/L

LL = 137–12.2 = 124.8 μmol/L

Note that the maximum permissible analytical variation (pCV_a) is:

$pCV_a < 0.5CV_I$

$pCV_a < 2.98\%$ or pSD = 4.1 μmol/L (at 137 μmol/L)

In this case the limits LL and UL are located, expressed in σ:

$12.2/4.1 = \pm 2.98\sigma$

The actual analytical performance in our laboratory is:

CV_a = 2.55%

SD_a = 3.5 μmol/L (137 μmol/L)

The Sigma metric is calculated according to this TAE model:

Sigma metric = $pTAE/CV_a$ = 8.87%/2.55% = 3.5

1.6. Comparison between the conventional Total error model, and the industrial model

When industrial quality concepts are compared with those in clinical chemistry, GageR&R-repeatability corresponds with imprecision. Between-operator variation in industry when handling the gage instrument corresponds with reproducibility.

In the case of monitoring, only the within subject variation is taken

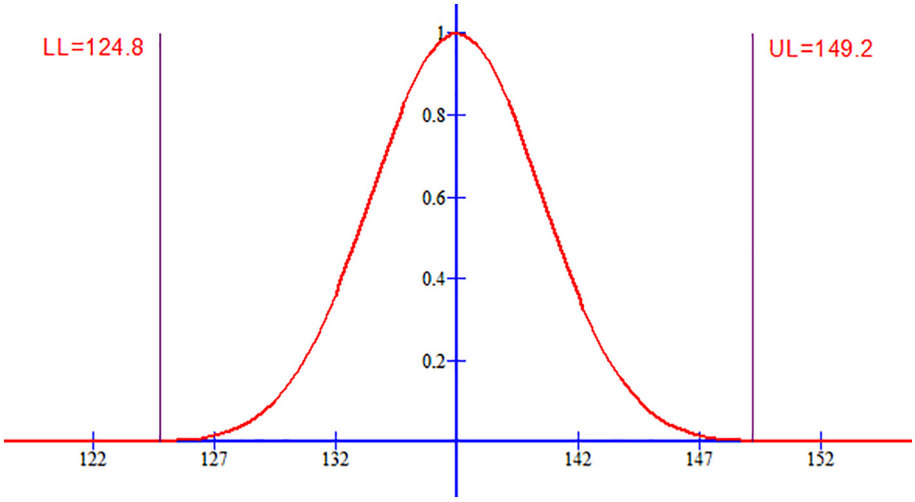

Fig. 3. Example for creatinine. The curve represents the actual analytical variation (SD_a = 3.5 μmol/L). Upper and lower performance limits (UL and LL) for standards are calculated applying the conventional total error model based on biological variation.

Table 2
* value from local laboratory as example. # Conventional TAE model [12].

Test	CV_a*	$CV_{intra\text{-}individual}$#	CV_{group}#	pTAE#	Sigma (TAE)	NDC
Creatinine	2.55%	5.95%	14.7%	8.9%	3.5	3.3
Sodium	1.11%	0.6%	0.7%	0.73%	0.66	0.76
Potassium	1.36%	4.6%	5.6%	5.6%	4.1	4.8
Glucose	0.70%	5.6%	7.5%	6.96%	9.9	11.3
Iron	1.77%	26.5%	23.2%	30.7%	17.3	21.2
Albumen	2.60%	3.2%	4.75%	4.07%	1.6	1.74
TSH	1.25%	19.3%	24.6%	38.2%	30.6	21.8

into account. If we exclude other pre-analytical sources of variation for simplicity, the expression for the Number of Distinct Categories transforms to:

$$NDC = \sqrt{2}\left(\frac{\sigma_{intra\ individual}}{\sigma_{measurement}}\right)$$

For the actual CV_a of creatinine (Fig. 4):

$$NDC = \sqrt{2}\left(\frac{CV_{intra\ individual}}{CV_{measurement}}\right) = \sqrt{2}\left(\frac{5.95}{2.55}\right) = 3.3$$

This should be compared with the minimum performance according the conventional models:

$$CV_{measurement} = 0.5\ CV_{intra\ individual} = 0.5\ (5.95) < 2.98\%$$

This would lead to a NDC of:

$$NDC = \sqrt{2}\left(\frac{CV_{intra\ individual}}{CV_{measurement}}\right) = \sqrt{2}\left(\frac{5.95}{2.98}\right) = 2.8$$

This means that this performance limit corresponds with a NDC value that is unacceptably low according to industry standards. Note that at a level of NDC = 4 the analytical variation would be $0.35CV_I$ instead of the conventional permissible analytical variation of $0.5CV_I$.

The concept of NDC can be related to the concept of the reference change value (RCV) (see note 2). Under certain assumptions, it can be shown that the performance limit of CV_a becomes: $CV_a < 0.35CV_I$.

2. Discussion

The conventional theory for the calculation of performance specifications is under debate, related to several flaws that are identified in the conventional model [2–4]. The two main flaws concern the addition of bias and imprecision terms, and the bias term. Both maxima of permissible bias and imprecision are added to obtain the permissible total analytical error (pTAE); these two maximum permissible errors are added, while these are derived under the mutually exclusive conditions of zero bias and zero imprecision, respectively. Secondly, the performance specification for imprecision (expressed as CV_a, coefficient of variation, analytical) is in general written as $CV_a < 0.5\ CV_I$. The biological within-subject variation CV_I is used here, based on the application of the test for monitoring. However, the maximum permissible bias was derived as $0.25CV_B$ or $0.25(CV_I^2 + CV_G^2)^{1/2}$, where the total biological variation CV_B includes both the within- and the between-

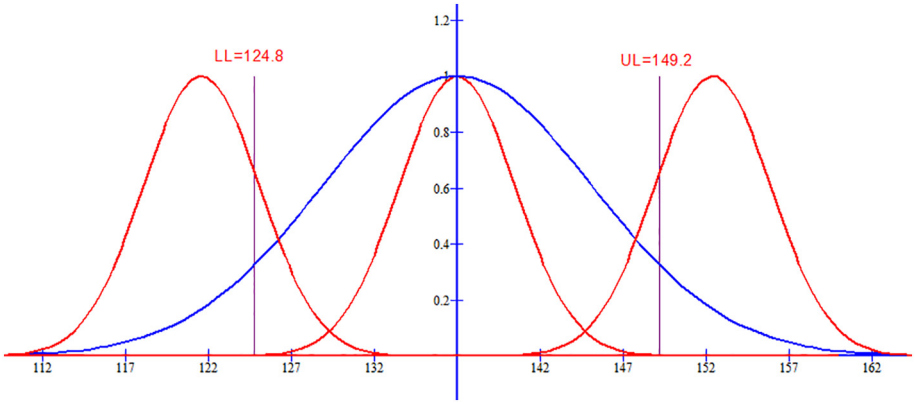

Fig. 4. The central curve represents the biological variation, the three curves the actual analytical variation with NDC 3.3.

subject variation (CV_G). The flaw identified here is, that this bias term is applied for performance specifications in case of monitoring although this expression had been derived from a reference interval model and applies to specifications for diagnosis where a higher pTAE is allowed.

We compared the models commonly applied in industry and the conventional total error model used in clinical chemistry. An important difference between both models is that the AIAG – although the Six-Sigma model is generally applied - assigns no sigma value to the quality of the measurement process, but only as a quality measure of the product and of the production process. Instead, the quality of the measuring procedure is expressed as a ratio of the product variation to short-time variation of the measurement procedure or NDC (Number of Distinct Categories). This concept is unknown in clinical chemistry.

In clinical chemistry, imprecision and bias are commonly distinguished as main measures of analytical error. In industry, more emphasis is put on the time factor that is related to the estimation of long term variation due to (slow) drifts in the measuring systems. This bias is not considered constant, but included in the model as long-term variation. Related to this, no linear combination of bias and imprecision is used as in the total error model and all factors are combined as variances and combined as root-square sum. The conditions including the time window under which bias is estimated should be defined in case of long-term bias (or drift). This time factor was also recognised in TE group, as this is an important complication in the definition of bias [2].

Gage R&R as a measure of short term measurement variability in industry can be compared with measurement variability estimated according to the CLSI EP15 protocol. Gage R&R and CLSI EP15 both measure combined repeatability and reproducibility. The analytical variation presented in Table 1 is the actual day-to-day variation and can be compared to analytical variation according to the CLSI protocol.

In the opinion of the authors this theoretical basis of the NDC-concept might be more solid than the theoretical basis of the maximum CV_a in clinical chemistry, where the limit of $0.5CV_I$ as the permissible quality level is related to an increase of the total variation of 12%. This increase is intuitively considered acceptably small. In contrast, the NDC is related to the concept of resolution in relation to the object to be measured. However, remarkably the analytical quality of some random laboratory tests, expressed as sigma metric and as NDC, gave results that were comparable.

We conclude that the calculation of the sigma metric is complicated due to a flawed pTAE model, and consensus is needed with regard to alternatives for analytical quality specifications. Models on measurement quality outside field of clinical chemistry apply other measures to express the quality of measurement that might lead to new insights.

Note 1.

In general:

$$\text{Sigma metric} = \left(\frac{UL - target}{\sigma_{measurement}}\right)$$

However, with the UL at $3\sigma_{part}$ from the target value, the sigma metric in the example (Fig. 2) would be:

$$\text{Sigma metric} = 3\left(\frac{\sigma_{part}}{\sigma_{measurement}}\right) = 3(2/0.7) = 8.6$$

In general, for the Sigma metric: with the UL at $N\sigma_{part}$ distance from the target

$$\text{Sigma metric} = NDC\left(\frac{N}{1.41}\right)$$

Obviously, there is an inconsistency here: the minimum quality is defined at NDC = 4. The minimum sigma-metric for acceptable quality is 3. With $N = 3$, and sigma = 3 the NDC is equal to 1.4, far below the minimum quality standard.

However, if we set NDC = 4 equal to 3 sigma:

Sigma = ¾NDC = >

$$¾\text{NDC} = ¾\sqrt{2}\left(\frac{\sigma_{part}}{\sigma_{measurement}}\right)$$

$$\left(2 = 3/(2\sqrt{2})\left(\frac{\sigma_{part}}{\sigma_{measurement}}\right)\right.$$

$$= 1.06\left(\frac{\sigma_{part}}{\sigma_{measurement}}\right)$$

Or:

$$\text{Sigma} \approx \left(\frac{\sigma_{part}}{\sigma_{measurement}}\right)$$

Note 2.

The concept of NDC can be related to the concept of the reference change value (RCV). The common expression for the RCV is [13]:

$$\text{RCV} = Z\sqrt{2}\ \sqrt{(CV_a^2 + CV_I^2)}$$

If we assume CV_a to be small compared to CV_I, this becomes:

$$\text{RCV} = Z\sqrt{2}\ CV_I \quad (1)$$

Rearranging:

$$CV_I = \frac{RCV}{Z\sqrt{2}} \quad (2)$$

The relation between NDC and RCV becomes (with Z = 2 and substituting CV_I in [3] with [2]):

$$\text{NDC} = \sqrt{2}\left(\frac{CV_I}{CV_a}\right) \quad (3)$$

$$\text{NDC} = \frac{RCV}{2\text{CV}_a}$$

If we assume the quality limit NDC ≥ 4

$$\frac{RCV}{CV_a} \geq 8$$

Substituting RCV with [1] (and with Z = 2) this becomes:

$$\frac{CV_I}{CV_a} = 4/\sqrt{2} = 2.8$$

Or:

$$\text{CV}_a = 0.35\text{CV}_I$$

This means that under the condition of NDC = 4, using the reference change model the analytical variation should be one-third or less of the within-subject variation.

References

[1] S. Sandberg, C.G. Fraser, A.R. Horvath, 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. 53 (2015) 833–835.

[2] W.P. Oosterhuis, H. Bayat, D. Armbruster, A. Coskun, K.P. Freeman, A. Kallner, D. Koch, F. Mackenzie, G. Migliarino, M. Orth, S. Sandberg, M.S. Sylte, S. Westgard, E. Theodorsson, The use of error and uncertainty methods in the medical laboratory, Clin. Chem. Lab. Med. 56 (2018) 209–219.

[3] W.P. Oosterhuis, Gross overestimation of total allowable error based on biological variation, Clin. Chem. 57 (2011) 1334–1336.

[4] W.P. Oosterhuis, S. Sandberg, Proposal for the modification of the conventional model for establishing performance specifications, Clin. Chem. Lab. Med. 53 (2015) 925–937.

[5] W.P. Oosterhuis, E. Theodorsson, Total error vs. measurement uncertainty: revolution or evolution? Clin. Chem. Lab. Med. 54 (2016) 235–239.

[6] D. Nevalainen, L. Berte, C. Kraft, E. Leigh, L. Picaso, T. Morgan, Evaluating laboratory performance on quality indicators with the six sigma scale, Arch. Pathol. Lab. Med. 124 (2000) 516–519.

[7] J.O. Westgard, Six Sigma Quality Design and Control: Desirable Precision and

Requisite QC for Laboratory Measurement Processes, Westgard QC, Inc., Madison, WI, 2001 (296 pp).
[8] G. Taguchi, Quality engineering (Taguchi methods) for the development of electronic circuit technology, IEEE Trans. Reliab. 44 (1995) 225–229.
[9] Measurement Systems Analysis Reference Manual, Fourth edition, Chrysler Group LLC, Ford Motor Company, General Motors Corporation (Automotive Industry Action Group, AIAG), Detroit-Michigan, USA, June 2010.
[10] C. Simion, A case study on Gage R&R in automotive industry, Ac J. Manufact. Eng. 13 (2015) 46–54.
[11] D. Stöckl, H. Baadenhuijsen, G. Callum, C.G. Fraser, J.C. Jean-Claude Libeer, P. Hyloft Petersen, et al., Desirable routine analytical goals for quantities assayed in serum, Eur. J. Clin. Chem. Clin. Biochem. 33 (1995) 157–169.
[12] C. Ricos, V. Alvarez, F. Cava, J.V. Garcia-Lario, A. Hernandez, C.V. Jimenez, J. Minchinela, C. Perich, M. Simon, Current databases on biologic variation: pros, cons and progress, Scand. J. Clin. Lab. Invest. 59 (1999) 491–500 (updated version see), https://www.westgard.com/biodatabase1.htm.
[13] C.G. Fraser, Reference change values, Clin. Chem. Lab. Med. 50 (2011) 807–812.