Opinion Paper

Elizabeta Topic*, Nora Nikolac, Mauro Panteghini, Elvar Theodorsson, Gian Luca Salvagno, Marijana Miler, Ana-Maria Simundic, Ilenia Infusino, Gunnar Nordin and Sten Westgard

How to assess the quality of your analytical method?

DOI 10.1515/cclm-2015-0869
Received for publication September 7, 2015

Abstract: Laboratory medicine is amongst the fastest growing fields in medicine, crucial in diagnosis, support of prevention and in the monitoring of disease for individual patients and for the evaluation of treatment for populations of patients. Therefore, high quality and safety in laboratory testing has a prominent role in high-quality healthcare. Applied knowledge and competencies of professionals in laboratory medicine increases the clinical value of laboratory results by decreasing laboratory errors, increasing appropriate utilization of tests, and increasing cost effectiveness. This collective paper provides insights into how to validate the laboratory assays and assess the quality of methods. It is a synopsis of the lectures at the 15th European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Continuing Postgraduate Course in Clinical Chemistry and Laboratory Medicine entitled "How to assess the quality of your method?" (Zagreb, Croatia, 24–25 October 2015). The leading topics to be discussed include who, what and when to do in validation/verification of methods, verification of imprecision and bias, verification of reference intervals, verification of qualitative test procedures, verification of blood collection systems, comparability of results among methods and analytical systems, limit of detection, limit of quantification and limit of decision, how to assess the measurement uncertainty, the optimal use of Internal Quality Control and External Quality Assessment data, Six Sigma metrics, performance specifications, as well as biological variation. This article, which continues the annual tradition of collective papers from the EFLM continuing postgraduate courses in clinical chemistry and laboratory medicine, aims to provide further contributions by discussing the quality of laboratory methods and measurements and, at the same time, to offer continuing professional development to the attendees.

Keywords: biological variation; detection limit; IQC/EQA; measurement uncertainty; method verification/validation; methods comparability.

***Corresponding author: Prof. Elizabeta Topic,** Chair Committee of Education and Training, Barutanski Jarak 35 A, 10000 Zagreb, Croatia, Phone: +385 91 3445 001, E-mail: elizabeta.topic@gmail.com
Nora Nikolac: University Department of Chemistry, Medical School University Hospital Sestre Milosrdnice, Zagreb, Croatia
Mauro Panteghini and Ilenia Infusino: Centre for Metrological Traceability in Laboratory Medicine (CIRME), University of Milan, Milan, Italy
Elvar Theodorsson: Department of Clinical Chemistry and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden
Gian Luca Salvagno: Laboratory of Clinical Biochemistry, Department of Neurological, Biomedical and Movement Sciences, University of Verona, Verona, Italy
Marijana Miler: University Department of Chemistry, Sestre Milosrdnice University Hospital Center, Zagreb, Croatia
Ana-Maria Simundic: University Department of Chemistry, Medical School University Hospital Sestre Milosrdnice, Zagreb, Croatia
Gunnar Nordin: Equalis, Uppsala, Sweden
Sten Westgard: Director Client Services and Technology Westgard QC, Orange, CT, USA

The EFLM Committee on Education and Training

Amongst the main missions of the EFLM is the education and training of its members. Through its Committee on Education and Training (C-ET), EFLM provides organized events in postgraduate continuous education in clinical chemistry and laboratory medicine. These activities started in 2001 in cooperation with Croatian Society of Medical Biochemistry and Laboratory Medicine (CSMBLM) and the Slovenian Association of Clinical Chemistry and Laboratory Medicine (SACCLM). For about 15 years, the C-ET has been providing attractive continuous education and training programs that are heterogeneous and diverse enough to meet the individual educational needs in the course of continuing professional development. The common title of EFLM courses is "New Classification,

Diagnosis and Treatment", each of them is dedicated to a particular medical discipline (Table 1). These EFLM Courses known as "Dubrovnik Courses" are affiliated with the programs of the Interuniversity Centre Dubrovnik.

This year the 15th EFLM course entitled "How to assess the quality of your method?" has moved from Dubrovnik to Zagreb to be more accessible to participants; in the future, it is anticipated that the courses will be held at different venues across Europe.

In addition to the Continuing Postgraduate Course, C-ET organizes other educational events including the "Symposium for Balkan region", which is commonly arranged in Belgrade, and the "Symposium on Education" held in Prague every 2 years.

Who, what and when to do in validation/verification of methods

Validation of a laboratory assay or method is defined as confirmation through the provision of objective evidence that the requirements for a specific intended use or application have been fulfilled. In-vitro diagnostics (IVD) manufacturers would be expected to provide such evidence as part of their design input [1]. Adequate method validation studies are needed before laboratory methods are considered for clinical use. The loop of the implementation design is indeed not closed until the finished IVD product is adequately validated to determine attributes and performance characteristics that meet the clinical needs. IVD manufacturers should define a calibration hierarchy to assign traceable values to their system calibrators and to fulfil, during this process, uncertainty limits for calibrators, which should represent a proportion of the uncertainty budget allowed for clinical laboratory results [2]. It is therefore important that, the laboratory profession clearly defines the clinically acceptable uncertainty for relevant tests [3] and end-users (i.e. clinical laboratories) may know and verify how manufacturers have implemented the traceability of their calibrators and estimated the corresponding uncertainty, including, if any, the employed goals [4]. Verification requires that there is sufficient objective evidence to determine that a given assay fulfils the specified requirements. In general, it should be possible to establish if the status of the measurement uncertainty budget associated with the proposed traceability chain is suitable or not for clinical application of the test [5]. Important tools for IVD traceability surveillance are the verification by clinical laboratories of the consistency of the declared performance during daily routine operations performed in accordance with the manufacturer's instructions and the organization of appropriately structured EQA programs. The former activity should be accomplished through the daily verification by clinical laboratories that control materials of analytical systems are in the manufacturer declared validation range (IQC component I) [6]. The participation to EQA schemes and meeting metrological criteria is mandatory. Target values for EQA materials (including their uncertainty) are optimally assigned with reference measurement procedures by accredited reference laboratories, these materials must be commutable and a clinically allowable inaccuracy for participant's results should be defined in order to prove the suitability of laboratory measurements in the clinical setting [7, 8]. Clinical laboratories should also separately monitor the imprecision of employed commercial systems through the IQC component II, primarily devoted to estimate the measurement uncertainty due to the random effects [2, 6].

Prior to method validation/verification, performance specifications for each measurement should be established.

Table 1: Years and topics discussed during EFLM Continuing Postgraduate Course in Clinical Chemistry and Laboratory Medicine.

Year	Topic[a]
2001	Diabetes mellitus
2002	Cardiovascular disease
2003	Neurological disease
2004	Neoplastic disease
2005	Autoimmune disease
2006	Metabolic syndrome
2007	Molecular diagnostics
2008	Kidney disease
2009	Thyroid disease
2010	Thrombophilia
2011	Inflammation
2012	Gastrointestinal disease
2013	Point-of-care testing
2014	Diabetes mellitus revisited
2015	Quality assessment of laboratory methods

[a]Available at: http://www.eflm.eu/index.php/educational-material.html.

Performance specifications

All test results are fraught with uncertainty despite every laboratories' ambition to its minimization. The knowledge of this uncertainty, observed during an extended period, is needed for the proper clinical use of the results. In order to compare uncertainty among different measurement

systems and methods, and to define performance specifications, we need tools to express the uncertainty and specify the performance numerically. Such data might also be used, e.g. to decide if it is possible to share common reference intervals and decision limits, or to decide if results from two assays are compatible [9].

A conference in 2014 arranged by the EFLM concluded that performance specifications should be based on one of three following models: clinical outcomes, biological variation or (if information from the first two sources is lacking) state-of-the-art of the assay performance [5].

For test primarily used for diagnostic purposes, the negative or positive predictive values in relation to some clinical outcome variable might be the most suitable way to define performance specifications. As negative and positive predictive values vary with the prevalence of the target disease, such performance specifications might differ depending on the setting in which the test is used.

Performance specifications based on biological variation have been proposed for more than 40 years. With the improvement in technology, the performance of assays usually improves. For instance, the devices for self-measurement of blood glucose perform today much better than 30 years ago, and consequently the quality requirements for the manufacturers have recently been revised [10]. For this reason, the best available technology (the "state-of-the-art") to a reasonable cost should always be encouraged.

A working group has recently been established within the EFLM (Task and Finish Group on Allocation of Laboratory Tests to Different Models for Performance Specifications – TFG-DM) in order to allocate the laboratory tests and the different use of them to these three different models for performance specifications. Possible criteria for allocation are: a) outcome model if the measurand has a central role in diagnosis and monitoring of a specific disease; b) biological variation model if the measurand has a high homeostatic control; and c) state-of-the-art model if neither central diagnostic role nor sufficient homeostatic control are shown.

When performance criteria for an assay are unmet, it is important to provide feedback to the manufacturer. If the laboratory profession agrees on common performance criteria, such feedback from users and organizers of EQA schemes will have a greater impact on the industry.

Biological variation

As stated previously, one of the ways to derive performance specifications rely on biological variability of the measurand. One of the most useful tools in recent years has been the development of the "Ricos' database", including specifications for desirable allowable total error, imprecision and bias, based on an ever-evolving literature review of biological variation of analytes [11, 12]. For many laboratories, the goals derived from biological variation represent the standards for quality performance. Yet, this use has not been without controversy: for instance, goals for some measurands are so demanding that no assay on the market can achieve them. Other goals are so wide that they are not demanding enough. The model with intra- and inter-individual variation is simple, but has also limitations. The sample-specific (matrix) error is not included in the model. The intra-individual variation is for several measurands literally "individual" and it might not be justified to base performance specifications on such an average variation, because the average is not representative for the majority of individuals. Finally, the estimated variation in healthy persons might not necessarily be representative for the variation observed in a diseased population [13].

To add to the controversy, the Milan 2014 EFLM conference called into question much of the validity of the biological variability information [5], and even went so far as to question the very accuracy of the underlying model for biologically-derived total allowable error [14].

How to assess the measurement uncertainty

The goal of standardization of measurements in laboratory medicine is to achieve compatible results in human samples, independent of the laboratory and/or the method used. This can be achieved by the adoption of the "reference system" approach, based on the concept of metrological traceability and a hierarchy of measurement procedures. The reference system requires reference procedures, reference materials and reference laboratories, which are able to produce results within defined limits of uncertainty [15]. The concept of uncertainty was introduced in the 1990s due to the lack of consensus on how to express the quality of measurement results. The International Vocabulary of Metrology (VIM) defines measurement uncertainty as a "non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used" [16]. Laboratorians may easily understand the meaning, but its estimate may be difficult in current

practice [3]. However, it must take into account that the estimation of measurement uncertainty is mandatory for reference measurement laboratories to obtain/maintain the accreditation according to ISO 17025:2003 and ISO 15195:2005 and for clinical laboratories to obtain the accreditation according to ISO 15189:2012 [17–19].

There are two approaches to estimate measurement uncertainty, the so-called "bottom-up" and "top-down" approaches. The "bottom-up" approach is the model proposed by the Guide to the Expression of Uncertainty of Measurement (GUM) based on a comprehensive dissection of the measurement, in which each potential source of uncertainty is identified, quantified and combined to generate a combined uncertainty of the result using statistical propagation rules [20]. This model has been fully endorsed by metrology institutions and suppliers of reference materials and is used in accredited laboratories that perform reference measurement procedures. The application of GUM in clinical laboratories is, however, not straightforward and has encountered many practical problems and objections [21]. As an alternative, the "top-down" approach described by Magnusson et al. can be used by clinical laboratories to estimate the measurement uncertainty of results, by using quality control data and certified reference materials for bias estimation [22, 23]. According to some experimental studies, uncertainties by "top-down" and "bottom-up" approaches, if correctly estimated, should be interchangeable.

Verification of imprecision and bias

A majority of the measurement methods used in laboratory medicine are produced by diagnostic companies, which have already validated them and established that they are fit for the intended purpose [4, 24]. The end-user laboratory, however, is requested to independently verify that the essential performance characteristics, including imprecision and bias of the measurement method and/or measurement system found during manufacturer's validation, can be reproduced locally. Verification is also required when substantial changes occur over time, e.g. change of a measurement system, relocation or when results of IQC or EQA schemes indicate that the performance of the method has worsened with time.

Local consensus on sufficient verification procedures have commonly been agreed and frequently influenced over time, e.g. by accreditation authorities. Published verification procedures have appeared rather recently [25–28]. The following is a brief summary of the most widely employed approaches:

- Bias studies. Clinical laboratories commonly measure in the order of 20–200 human samples having as wide a concentration range as possible, using both the comparison ("reference") method and the evaluated method. At least 20 repeated measurements of at least two pooled patient samples may also be used. This latter approach may actually be an advantage when the medical decision limit is close to the detection limit of the measurement method or system.
- Imprecision studies. For estimating imprecision, suitable stable control materials for IQC at two concentration levels are measured in at least two replicates for at least 5 consecutive days each week for 2 weeks.
- Data presentation and analysis. Linear regression, preferably orthogonal linear regression [29, 30], bias plots [31, 32] and analysis of variance [33] techniques are used to quantify bias and within- and between-series imprecision, respectively.

Limit of blank, limit of detection, limit of quantification and limit of decision

Limit of blank (LoB), limit of detection (LOD), limit of quantitation (LOQ) and limit of decision are concepts and terms used to describe the lowest concentration of a measurand that can be reliably measured by a measurement procedure [16, 34–36]. The literature in this area has previously been and is unfortunately still confusing regarding concepts, nomenclature and methods. The approach recommended here is primarily based on recent recommendations by Eurachem [34].

- The LoB is the highest apparent concentration of a measurand expected when replicates of a blank sample containing no measurand are measured. The LoB refers to test results or signals and not to actual concentrations.
- The limit of decision (CCα) is the concentration of the measurand that is significantly different from zero. The concept is, e.g. used when determining whether a material is contaminated or not.
- LOD is the lowest concentration of the measurand detectable at a specified level of confidence. The LOD of the measurement system/instrument and of the method should be kept apart. The LOD of

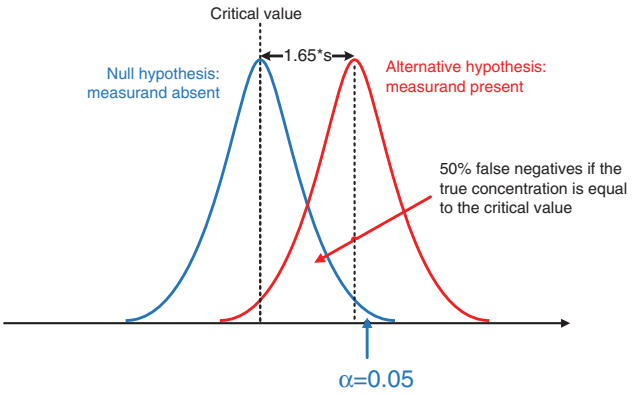

Figure 1: Neyman–Pearson theory concept.

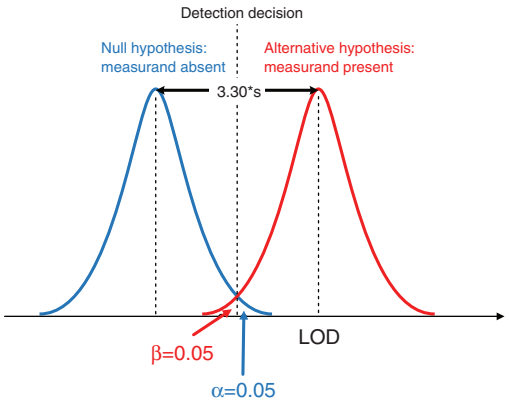

Figure 2: Estimation of limit of detection.

the measurement system is determined by presenting the system directly with the reagent blank or with other types of samples. When the LOD of the measurement method is determined, the sample is processed through all the steps of the measurement procedure.

– LOQ is the lowest concentration at which the performance of a method or measurement system is acceptable for a specified use.

The Neyman-Pearson theory [37] provides methods for calculating probabilities when choosing between two alternative hypotheses (Figure 1). This theory is important for the current understanding of determining, e.g. LOD and LOQ. Estimating the LOD means, e.g. choosing false-positive probability of α=0.05, which leads to a critical value of approximately 1.65*s* (where s is the standard deviation of a large number of results for a blank sample or a sample containing a low concentration of the measurand, and 1.65 is the one-tailed Student's t-value for infinite degrees of freedom at the significance level α=0.05). In order to avoid too high false-negative measurement results, the false-negative error also needs to be appropriately set, commonly $\beta=\alpha$=0.05. Calculating the LOD with $\alpha=\beta$=0.05 will therefore be 1.65+1.65=3.30, which is frequently rounded to 3*s*' (Figure 2).

For a statistically proper estimate of the LOD, the multiplying factor used should take into account the number of degrees of freedom associated with the estimate of *s*. For example, if s is obtained from 10 replicate measurements, the Student's t-value at α=0.05 is 1.83 (9 degrees of freedom). This leads to an LOD calculated as 3.7*s* (Figure 3).

Tables 2 and 3 present strategies for calculating LOD and LOQ, respectively. The calculation of the LOQ as described in the Table 3 is appropriate when detecting very low concentrations of measurands, e.g. in environmental analysis or when detecting drugs of abuse. In this situation, a pragmatic approach defining LOQ as equal to the lowest concentration at which the CV is ≤5% appears to be appropriate [38, 39].

Analytical specifications in clinical laboratories are dictated by the clinical use of the measurement methods. Many clinical laboratories therefore prefer to apply other definitions of LOQ than the ones commonly used in international metrology. It is therefore crucial to specify which definition is used when reporting LOQ.

Statistical approaches to compare methods and analytical systems

When a new analytical system or method is replacing the existing one, laboratory professionals have to investigate if there are differences between obtained results that could have an impact on clinical decision-making. Thus, result equivalence should be checked, even if, e.g. the exact same model of analyzer is introduced.

There are several statistical approaches to investigate method comparability and most of them can be appropriate. Unfortunately, we often witness the misuse of statistics, especially in the manufacturer's declarations. IVD manufacturers often substantiate the claim that two methods are comparable by a high correlation coefficient (e.g. 0.99). However, the use of the correlation coefficient is not adequate for showing result equivalence, since a high correlation only tells us that the two sets of data are highly related [40]. Passing and Bablok (P-B) regression analysis gives us more information and enables to determine if constant or proportional difference is present between the methods. In order to evaluate statistical

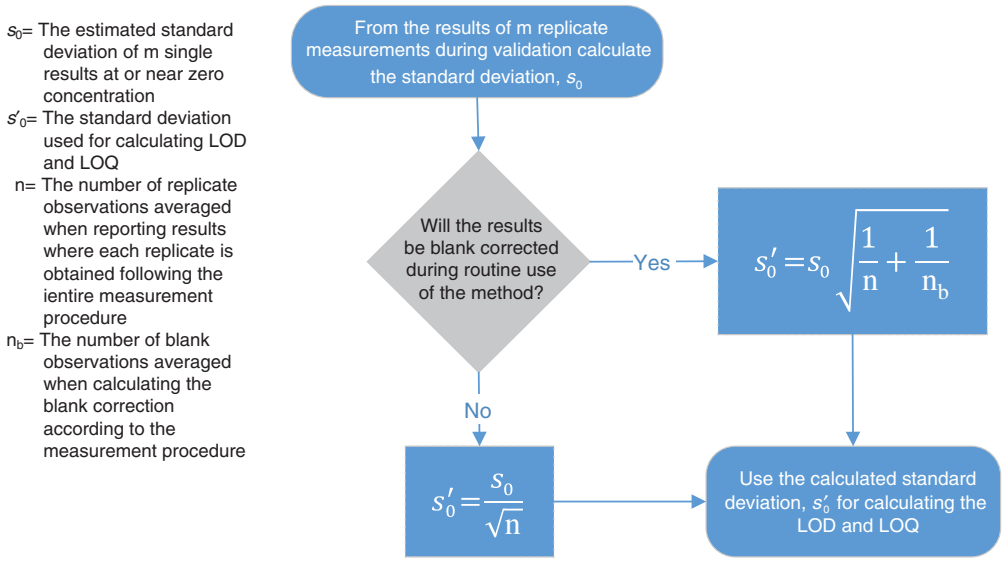

Figure 3: How to calculate LOD and LOQ [34].

significance of the intercept (from zero) and slope (from the unity), data has to be presented using 95% confidence intervals (CI) [41]. If data on CI are missing, one can wrongly interpret results of regression analysis. This lack is often found in reagent package inserts. P-B analysis has, however, some drawbacks, as different sets of data can show exactly the same regression equation. In addition, analysis of residuals is required to determine the amount of data that can be explained with the model [42]. Deming regression is another type of regression analysis, which takes into account the analytical variability of both tested methods (derived from duplicate measurements or method CV).

Regression analysis does not provide the information about differences between specific pairs of measurement. The Bland-Altman (bias) plot is the best approach to evaluate the differences between methods. In this difference plot, the mean between two methods is presented on the x-axis (unless the comparison method is considered as the "reference"). The choice of variable on the y-axis depends on the type of the difference between methods: for checking a constant bias, it is better to present the absolute difference between methods, while percentage difference is better suited for evaluating proportional bias. Limits of agreement define borders within which 95% of differences. Overall

Table 2: Calculation of limit of detection [34].

What to do	How many times	What to calculate from the data	Comments
a) Replicate measurements of blank samples, i.e. matrices containing no measurand and replicate measurement of test samples with low concentrations of the measurand	10	Calculate the standard deviation, s_0 of the results Calculate s'_0 from s_0 as shown in Figure 3 Calculate LOD=3*s'_0	
b) Replicate measurements of reagent blanks and replicate measurements reagent blanks spiked with low concentrations of measurand[a]	10	Calculate the standard deviation, s_0 of the results Calculate s'_0 from s_0 as shown in Figure 1 Calculate LOD=3*s'_0	Approach b) is acceptable, when it is not possible to obtain blank samples or test samples at low concentrations When these blanks do not go through the whole measurement procedure the calculation will give instrumental LOD

[a]Spiking can compromise the commutability of the sample.

Table 3: Calculation of limit of quantitation [34].

What to do	How many times	What to calculate from the data	Comments
a) Replicate measurements of blank samples, i.e. matrices containing no detectable measurand or replicate measurements of test samples with low concentrations of analyte	10	Calculate s_0' from s_0 as shown in Figure 3 Calculate LOQ as $LOQ=k_Q \times s_0'$	The value for the multiplier k_Q is usually 10, but other values such as five or six are commonly used (based on "fitness for purpose" criteria)
b) Replicate measurements of reagent blanks or replicate measurements of reagent blanks spiked with low concentrations of measurand	10	Calculate the standard deviation, s_0 of the results Calculate s_0' from s_0 as shown in Figure 3 Calculate LOQ as $LOQ=k_Q \times s_0'$	Approach b) is acceptable, when it is not possible to obtain blank samples or test samples at low concentrations When these reagent blanks are not taken through the whole measurement procedure and are presented directly to the instrument the calculation will give the instrument LOQ

1) For some measurement systems, e.g. chromatography, a test sample containing too low a concentration or a reagent blank might need to be spiked in order to get a non-zero standard deviation
2) The entire measurement procedure should be repeated for each determination
3) The standard deviation is expressed in concentration units

The application of this calculation should be discussed more extensively for its impact on the clinical application of the measurement. Namely, in the selection of the multiplier k it appears subjective and not based on objectively derived criteria. In a more practical way, WHO-ECBS defined LOQ as the lowest amount of measurand that can be quantitatively determined with stated acceptable imprecision and bias. It means that LOQ should be defined by that concentration fulfilling analytical goals to make the measurement clinically meaningful.

bias can be evaluated by constructing the line of equity and 95% CI (if there is no bias, the line of equity corresponds to zero) [43].

Statistical tests can determine if the bias between methods is significant. This, however, tells us nothing about the clinical significance of the difference. In order to evaluate the latter, acceptability criteria derived by the models listed in Table 4 should be applied [5]. Only if the determined bias values exceed established specifications, can we conclude that there is a clinically relevant difference between methods. Laboratories should inform clinicians about the issue and the deriving effect it can have on the interpretation of the patient's result.

Verification of reference intervals

Accreditation programs play a pivotal role in clinical laboratories for the management of the patient safety. The latest (third) revision of the ISO 15189 in 2012 [17] emphasizes that "biological reference intervals shall be periodically reviewed" by laboratory personnel and that they should be verified every time a variation in analytical and/or pre-analytical procedures occurs [44]. This requirement poses challenges to laboratory personnel, considering the large number and different types of clinical laboratory tests, as well as the fast development of analytical technology [45]. The directive of the European Union on in

Table 4: Recommended models for defining analytical performance specifications.

Recommended models for defining analytical performance specifications.
Model 1: Based on the effect of analytical performance on clinical outcomes
a. Done by direct outcome studies – investigating the impact of analytical performance of the test on clinical outcomes;
b. Done by indirect outcome studies – investigating the impact of analytical performance of the test on clinical classifications or decisions and thereby on the probability of patient outcomes, e.g. by simulation or decision analysis
Model 2: Based on components of biological variation of the measurand
Model 3: Based on state of the art of the measurement, defined as the highest level of analytical performance technically achievable

Adapted from Sandberg et al. [5].

vitro diagnostic medical devices (Directive 98/79/EC) [22] states that IVD manufacturers need to provide "detailed information on reference intervals for the quantities being determined, including a description of the appropriate reference population" [46].

In clinical practice, a widespread and practiced way for interpreting laboratory results rely on a two-sided comparison based on reference intervals [45]. However, at the dawn of the 21st century, there is now a defined priority to implement in quality system of clinical laboratories a specific procedure for establishing, verifying and revising reference values [47]. Gräsbeck and Saris first introduced the concept of reference values in 1969 [48]. Nevertheless, the correct approach for definition, implementation and use of reference intervals remains a critical issue in laboratory medicine. As defined by the IFCC [49], and recently reviewed by the Clinical and Laboratory Standards Institute (CLSI), the term 'reference interval' entails a range of values obtained from individuals appropriately selected in order to satisfy suitably defined criteria [50, 51]. The clinical laboratory staff has to define and consistently verify the accuracy of pre-analytical conditions, the analytical method and its performance and the characteristics of the population to be analyzed [43]. The main preconditions to be addressed for defining a reference interval in ostensibly healthy subjects are listed in Table 5.

Once established, reference intervals should be locally validated. As recently discussed [52], the validation can be done according to the CLSI document C28-A3, paragraph 11.2, by examining 20 reference individuals from a laboratory's own subject population. If no more than two (10%) of the 20 tested values fall outside the previously established reference interval, this can be locally adopted.

Verification of qualitative test procedures

The qualitative (ordinal scale) laboratory methods can be used for screening, diagnosis or monitoring of disease and treatment response. In general, qualitative methods have only two possible results: "positive" and "negative". Some qualitative tests derived from dichotomized quantitative tests are sometimes semi-quantitative. Examples of qualitative tests in laboratories are immunology screening tests (done by immunofluorescence), some molecular tests and urinalysis using urine test strips. According to the ISO 15189:2012 standard, the verification for all types of methods should be performed before their implementation in the laboratory work [19]. The protocol for the verification can be defined by the laboratory and it may not be the same for every user [53]. However, all of tests should meet predefined performance characteristics and provide reproducible and accurate results [54].

As recommended by the CLSI, verification of the qualitative methods should include studies on imprecision, bias and method comparison [55]. Particularly, the trueness of the method should be verified if quantitative methods are proposed for concentration measurements. The reproducibility for the analyte measurements near the cut-off concentrations should also be performed, mainly if results are derived from quantitative values and reported binary as positive/negative. The use of positive and negative samples at concentrations 20% lower and higher than cut-off values is recommended [56]. If high positive or low negative samples would indeed be used the problem with results in the area of clinical decision cannot be detected.

A method comparison study has to be performed between the comparison method in use (considered as the "reference" method) and the new qualitative test procedure. Comparison of two qualitative methods or between qualitative and quantitative methods has specific rules and recommended statistical analyses used in quantitative method comparison, such as Bland-Altman plots or P-B regression, cannot be used. Results from method comparison should be shown in 2×2 table as the ratio of agreement between the new method and the quantitative test or diagnosis adjudication, if available. From that table, the ratio of true positive and true negative values or related diagnostic sensitivity and specificity can be calculated.

Table 5: Conditions to be addressed for correctly defining reference intervals.

Conditions to be addressed for correctly defining reference intervals
– Definition of the basic demographic characteristics of reference groups of individuals;
– Pre-analytical and analytical criteria should be fulfilled;
– Results should be obtained using a standardized (advisably traceable) method, in a system with defined analytical specifications and by employing an appropriate quality control program;
– The diagnostic characteristics (e.g. sensitivity, specificity, predictive values) of the assay should be known in advance [49]
– The statistical analysis for evaluating value distributions and deriving intervals should be based on appropriate tests

As described above for quantitative methods, reference intervals should be checked. Samples for verification of reference values should be 20% lower and higher than cut-off value.

The number of samples should be predefined regarding the specific verification part. For reproducibility testing, a minimum of 20 samples should be used and weighted kappa coefficients should be calculated [57]. The Kappa coefficient can be calculated with at least 30 samples, a minimum of 10 samples from each category (positive and negative).

Results of qualitative methods verification should be interpreted according to the predefined acceptance criteria. The new method can be implemented as the part of clinical laboratory procedures when these criteria are fulfilled.

How to best use your IQC and EQA data

Traditionally, IQC uses sample materials with assigned values and IQC results are evaluated continuously in relation to these known values. Although the use is "internal", the outcome can be compared with results obtained from other laboratories, using the same materials and devices.

EQA schemes should be used to evaluate trueness and accuracy of laboratory assays. The EQA material ought therefore to be commutable and it is important for the EQA organizer to assign values for the measurands in the material as close as possible to true values [58, 59]. The ultimate way to assign values is by using reference measurement procedures. However, the availability of reference measurement procedures is limited and the cost is relatively high. Therefore, other ways could be used to assign values. One such substitute is the transfer of certified reference values from reference materials to EQA materials by measurements in parallel. Again, a prerequisite for such procedure is that both the reference material and the EQA material are commutable.

For the most common measurands, it is expected that the major CE-marked IVD products today produce measurement results close to reference measurement results. Consequently, some authors have proposed that, although the different method group mean values might vary slightly randomly over time, it is reasonable to assume that a mean value of the method group mean values is close to a true value.

A consensus-based grading (e.g. "grand mean", method- or commercial system-specific mean/median value) is often used when the reference value is lacking. This procedure to assign value suffers from a limitation that the most common analytical systems in the market contribute with a higher weight to the mean value. Furthermore, if measurement methods that contribute with results in the pool not are fully documented, the traceability of the consensus value may be questionable [60].

The EQA participant results should be evaluated against agreed limits. These limits (recently discussed in Milan at the EFLM conference [61]) have been agreed by either professional organizations, authorities or suggested by the EQA organizers. The participants might also create their own acceptance limits, with respect to a specific use of a test.

The reason for deviating results must always be searched for. The most common reason for a deviating result might be a transfer mistake. Ideally, it is a recommended to review EQA data according to a structured scheme [62].

Ordinal scale results cannot be evaluated in the same way as quantitative results. For variables with an underlying quantitative scale, the "true" or assigned value can be established with a reference method or well-controlled measurement procedures. Close to the equivalence point, or c50-value, both "negative" and "positive" results are expected. The information from such a survey is therefore the location of the c50 value for the assay. Use of materials with expected values that are very different from the c50 value is recommended to evaluate the performance of the individual user of the test [63].

Nominal scale tests are tests were the value of the "examinand" is identified or named. Examples of these tests are recognition of cell types, bacteria species or blood types. In this case, the results in EQA schemes are evaluated as, e.g. a fraction of correctly identified objects [64].

Six-Sigma metrics

At the end of method validation or verification, the laboratory has collected data, crunched numbers, and created some graphs. Now what? With all the different studies and statistics, how can we synthesize the results into a single verdict: acceptable or unacceptable? Is a method with high bias but low imprecision or a method with low bias but high imprecision acceptable?

The sigma metric approach allows a laboratory to take a broader view of the data, putting together estimates of bias and imprecision into a practical judgment on the clinical usefulness of the method. With sigma metrics, laboratories can not only determine whether or not the method is "good enough" for patient care, but they can also estimate the number of defects that will be generated by the assay,

as well as apply the appropriate quality control specifications to help assure the quality of the results will always match the needs of the patient.

Verification of blood collection system

Clinical laboratories have been at the vortex of the maelstrom affecting medicine over the past few years. Various approaches are being implemented to reduce overall expenditure for laboratory services, such as centralization and consolidation of facilities, increasing level of automation, decentralizing testing (i.e. point-of-care testing). In many of these situations, most problems in the pre-analytical phase entail factors directly associated with blood specimen collection.

In the past, the lack of standardized procedures for sample collection accounted for most of the mistakes encountered within the total examination process [65]. Data a which emerged from representative studies showed that problems directly related to specimen collection were the main source of diagnostic errors and variability, including hemolyzed, insufficient, clotted, and incorrect blood samples [65]. These problems related to inappropriate procedures for collection and handling of specimens, i.e. use of improper collection tools, prolonged stasis during venepuncture, time before centrifugation or analysis, unsuitable storage, etc. [66]. Therefore, the choice of devices for blood collection becomes a pivotal aspect in optimizing the pre-analytical phase and achieving reliable testing results [67]. Moreover, accreditation programs for medical laboratories emphasize that the laboratory personnel need to evaluate the influence of blood collection systems on analytical quality and estimate the measurement uncertainty (ISO 15189:2012) [19].

For evaluation of blood collection systems, clear indications have been provided by the CLSI guidelines on collection of blood specimens for laboratory testing and protocol to evaluate the different type of commercial blood collection needles and tubes [68–70].

A large number of venous blood collection system (including needles, tubes, etc.) are currently commercially available. Venepunctures have traditionally been carried out using ordinary straight or butterfly needles. Laboratories need to verify the influence of the needle used and blood drawing technique [67]. In order to secure reliability of test results, CLSI has also defined a protocol that manufacturers should follow for tube validation, which is similar to protocols used for laboratory test validation [66]. Additional indications have been published at the national level, specifically aimed to ensure that blood collection systems fulfil specific requisites of quality, workability and efficiency [71]. It is also noteworthy that a number of pre-analytical mistakes are attributable to insufficient audits with healthcare operators involved in specimen collection/handling.

Standardization and monitoring of all pre-analytical variables would be associated with the best organizational and clinical outcomes. The governance of the entire examination process (thus including the evaluation of blood collection system) will also reduce laboratory costs and enhance clinician-laboratory cooperation.

In summary, in agreement with the aim of the EFLM Continuing Postgraduate Course on method validation and verification held in Zagreb, the main scope of this collective paper was to enable the exchange of ideas and knowledge related to the most common issues and everyday problems found in clinical laboratories, in order to ensure a better quality of daily laboratory results.

Author contributions: All the authors have 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.
Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Powers DM, Greenberg N. Development and use of analytical quality specifications in the in vitro diagnostics medical device industry. Scand J Clin Lab Invest 1999;59:539–44.
2. Braga F, Infusino I, Panteghini M. Performance criteria for combined uncertainty budget in the implementation of metrological traceability. Clin Chem Lab Med 2015;53:905–12.
3. Panteghini M. Application of traceability concepts to analytical quality control may reconcile total error with uncertainty of measurement. Clin Chem Lab Med 2010;48:7–10.
4. Braga F, Panteghini M. Verification of in vitro medical diagnostics (IVD) metrological traceability: responsibilities and strategies. Clin Chim Acta 2014;432:55–61.
5. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, 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 2015;53:833–5.

6. Braga F, Infusino I, Panteghini M. Role and responsibilities of laboratory medicine specialists in the verification of metrological traceability of in vitro medical diagnostics. J Med Biochem 2015;34:282–87.
7. Infusino I, Schumann G, Ceriotti F, Panteghini M. Standardization in clinical enzymology: a challenge for the theory of metrological traceability. Clin Chem Lab Med 2010;48:301–7.
8. Braga F, Panteghini M. Standardization and analytical goals for glycated hemoglobin measurement. Clin Chem Lab Med 2013;51:1719–26.
9. Hyltoft Petersen P, Jensen EA, Brandslund I. Analytical performance, reference values and decision limits. A need to differentiate between reference intervals and decision limits and to defi ne analytical quality specifications. Clin Chem Lab Med 2012;50:819–31.
10. ISO. In vitro diagnostic test systems – Requirements for blood-glucose monitoring systems for self-testing in managing diabetes mellitus. ISO 15197:2013, 2nd ed. Geneva: ISO, 2015.
11. Ricos C, Alvarez V, Cava F, Garcia-Lario JV, Hernandez A, Jimenez CV, et al. Current databases on biologic variation: pros, cons and progress. Scand J Clin Lab Invest 1999;59:491–500. This database was most recently updated in 2014.
12. Perich C, Minchinela J, Ricos C, Fernández-Calle P, Alvarez V, Doménech MV, et al. Biological variation database: structure and criteria used for generation and update. Clin Chem Lab Med 2015;53:299–305.
13. Bartlett WA, Braga F, Carobene A, Coskun A, Prusa R, Fernandez-Calle P, et al. A checklist for critical appraisal of studies of biological variation. Clin Chem Lab Med 2015;53:879–85.
14. Oosterhuis WP. Gross overestimation of total allowable error based on biological variation. Clin Chem 2011;57:1334–6.
15. Panteghini M. Traceability as a unique tool to improve standardization in laboratory medicine. Clinical Biochemistry 2009;42:236–40.
16. JCGM. Evaluation of measurement data. Guide to the expression of uncertainty in measurement. JCGM 100:2008, GUM 1995 with minor corrections. Available at: http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf: Joint Committee for Guides in Metrology; 2008. Accessed June 30, 2015.
17. ISO. General requirements for the competence of testing and calibration laboratories. ISO 17025. Geneva: ISO, 2005.
18. ISO. Laboratory medicine – Requirements for reference measurements laboratories. ISO 15195. Geneva: ISO, 2003.
19. ISO. Medical laboratories – Particular requirements for quality and competence ISO 15189. Geneva: ISO, 2012.
20. JCGM 100:2008 Evaluation of measurement data – Guide to the expression of uncertainty in measurement.
21. Lee JH, Choi J-H, Youn JS, Cha YJ, Song W, Park AJ. Comparison between bottom-up and top-down approaches in the estimation of measurement uncertainty. Clin Chem Lab Med 2015;53:1025–32.
22. Magnusson B, Naykki T, Hovind H, Krysell M. Handbook for Calculation of Measurement Uncertainty in environmental laboratories. NORDTEST Report TR 537 – 2003-05.
23. Ceriotti F, Brugnoni D, Mattioli S. How to define a significant deviation from the expected internal quality control result. Clin Chem Lab Med 2015;53:913–8.
24. EU. Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on in vitro diagnostic medical devices. Eur-Lex. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31998L0079:EN:NOT; 1998. Accessed June 30, 2015.
25. Burnett D. Measurement verification in the clinical laboratory: A guide to assessing analytical performance during the acceptance testing of methods (quantitative examination procedures) and/or analysers. London: The Association for Clinical Biochemistry. Available at: http://www.acb.org.uk/An%20Ver/David%20Burnett%20Editorial.pdf, 2010. Accessed June 30, 2015.
26. Khatami Z, Hill R, Sturgeon C, Kearney E, Breadon P, Kallner A. Measurement verification in the clinical laboratory: A guide to assessing analytical performance during the acceptance testing of methods (quantitative examination procedures) and/or analysers. London: The Scientific Committee of the Association for Clinical Biochemistry. Available at: http://www.acb.org.uk/An%20Ver/Measurement_ver_09.27.pdf, 2010. Accessed June 30, 2015.
27. Theodorsson E. Validation and verification of measurement methods in clinical chemistry. Bioanalysis 2012;4:305–20.
28. CLSI. User Verification of Performance for Precision and Trueness; Approved Guideline EP15-A2: Clinical and Laboratory Standards Institute; 2006.
29. Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. Application of linear regression procedures for method comparison studies in clinical chemistry. Part II. J Clin Chem Clin Biochem 1984;22:431–45.
30. Deming WE. Statistical adjustment of data. New York, USA: John Wiley & Sons, Inc., 1943.
31. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.
32. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. The Statistician 1983;32:307–17.
33. Kallner A. Laboratory statistics: handbook of formulas and terms, 1st ed. Amsterdam: Elsevier, 2014:xiv, 139.
34. Magnusson B, Örnemark U. Eurachem Guide: The Fitness for Purpose of Analytical Methods – A Laboratory Guide to Method Validation and Related Topics. Available from www.eurachem.org. Eurachem, 2014.
35. JCGM. International vocabulary of metrology – Basic and general concepts and associated terms (VIM 3): Bureau International des Poids et Mesures, 3rd ed. 2012. Available at: http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf. Accessed June 30, 2015.
36. Ellison SL, Farrant TJ, Barwick V. Royal Society of Chemistry (Great Britain). Practical statistics for the analytical scientist: a bench guide, 2nd ed. Cambridge, UK: RSC Publishing, 2009:xiv, 268.
37. Lehmann EL. Fisher, Neyman, and the creation of classical statistics. New York, NY: Springer, 2011:viii, 115.
38. Coleman D, Auses J, Grams N. Regulation – from an industry perspective or relationships between detection limits, quantitation limits, and significant digits. Chemom Intell Lab Syst 1997;37:71–80.
39. Jill Carlson J, Artur Wysoczanski A, Edward Voigtman E. Limits of quantitation – Yet another suggestion. Spectrochim Acta B 2014;96:69–73.
40. Udovičić M, Baždarić K, Bilić-Zulle L, Petrovečki M. What we need to know when calculating the coefficient of correlation? Biochem Med (Zagreb) 2007;17:10–5.

41. Simundic AM. Confidence interval. Biochem Med (Zagreb) 2008;18:154–61.
42. Bilić-Zulle L. Comparison of methods: Passing and Bablok regression. Biochem Med (Zagreb) 2011;21:49–52.
43. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb) 2015;25:141–51.
44. Plebani M, Lippi G. Reference values and the journal: why the past is now present. Clin Chem Lab Med 2012;50:761–3.
45. Ceriotti F, Hinzmann R, Panteghini M. Reference intervals: the way forward. Ann Clin Biochem 2009;46(Pt 1):8–17.
46. Ceriotti F. Prerequisites for use of common reference intervals. Clin Biochem Rev 2007;28:115–21.
47. Siest G, Henny J, Gräsbeck R, Wilding P, Petitclerc C, Queraltó JM, et al. The theory of reference values: an unfinished symphony. Clin Chem Lab Med 2013;51:47–64.
48. Grasbeck R, Saris NE. Establishment and use of normal values. Scand J Clin Lab Invest 1969;26(Suppl 110):62–3.
49. Guidi GC, Salvagno GL. Reference intervals as a tool for total quality management. Biochem Med (Zagreb) 2010;20:165–72.
50. Solberg HE. A Guide to IFCC recommendations on reference values. JIFCC 1993;5:160–4.
51. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline – third edition CLSI document C28-A3. Wayne, PA: Clinical and Laboratory Standards Institute, 2008.
52. Panteghini M, Ceriotti F. Obtaining reference intervals traceable to reference measurement systems: it is possible, who is responsible, what is the strategy? Clin Chem Lab Med 2012;50:813–7.
53. Burd EM. Validation of Laboratory-Developed Molecular Assays for Infectious Diseases. Clin Microbiol Rev 2010;23:550–76.
54. Nichols JH. Verification of method performance for clinical laboratories. Adv Clin Chem 2009;47:121–37.
55. CLSI. User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline – Second Edition. CLSI document EP12-A2. Wayne, PA: Clinical and Laboratory Standards Institute, 2008.
56. Nordin G. Before defining performance criteria we must agree on what a "qualitative test procedure" is. Clin Chem Lab Med 2015;53:939–41.
57. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276–82.
58. Miller WG, Myers GL, Rej R. Why commutability matters. Clin Chem 2006;52:553–4.
59. Korzun WJ, Nilsson G, Bachmann LM, Myers GL, Sakurabayashi I, Nakajima K, et al. Difference in Bias Approach for Commutability Assessment: Application to Frozen Pools of Human Serum Measured by 8 Direct Methods for HDL and LDL Cholesterol. Clin Chem 2015;61:1107–13.
60. De Bièvre P. Is "consensus value" a correct term for the product of pooling measurement results? Accred Qual Assur 2012;17:639–40.
61. Dallas Jones GR. Analytical performance specifications for EQA schemes – need for harmonisation. Clin Chem Lab Med 2015;53:919–24.
62. Rustad P, Kristensen GB. Errors in an EQA result – what is the cause? Labquality News 2013:18–20.
63. Hyltoft Petersen P, Christensen NG, Sandberg S, Nordin G, Pedersen M. How to deal with dichotomous tests? Application of a rankit ordinal scale model with examples from the Nordic ordinal scale project on screening tests. Scand J Clin Lab Invest 2008;68:298–311.
64. Restelli V, Bhuvanendran S, Lee C, Kwok E, Noble M. Performance of Canadian clinical laboratories processing throat culture proficiency testing surveys. Accred Qual Assur 2014;19:445–50.
65. Lippi G, Guidi GC, Mattiuzzi C, Plebani M. Preanalytical variability: the dark side of the moon in laboratory testing. Clin Chem Lab Med 2006;44:358–65.
66. Lippi G, Becan-McBride K, Behúlová D, Bowen RA, Church S, Delanghe J, et al. Preanalytical quality improvement: in quality we trust. Clin Chem Lab Med 2013;51:229–41.
67. Lippi G, Salvagno GL, Brocco G, Guidi GC. Preanalytical variability in laboratory testing: influence of the blood drawing technique. Clin Chem Lab Med 2005;43:319–25.
68. CLSI. Procedures for collection of diagnostic blood specimens by venipuncture; approved guideline, 6th ed. CLSI document H3-A6. CLSI: Wayne, PA: Clinical and Laboratory Standards Institute, 2007.
69. CLSI. Specimen labels: Content and location, fonts, and label orientation; approved standard. CLSI document AUTO12-A. Wayne, PA: Clinical and Laboratory Standards Institute, 2011.
70. CLSI. Validation and Verification of Tubes for Venous and Capillary Blood Specimen Collection; Approved Guideline. CLSI document GP34-A. Wayne, PA: Clinical and Laboratory Standards Institute, 2010.
71. Plebani M, Caputo M, Giavarina D, Lippi G. Methodological notes on acquisition and use of close evacuated systems for collection, handling and storage of venous blood samples for laboratory diagnostics. Biochim Clin 2013;37:303–11.