Opinion Paper

Wytze P. Oosterhuis*, Hassan Bayat, David Armbruster, Abdurrahman Coskun, Kathleen P. Freeman, Anders Kallner, David Koch, Finlay Mackenzie, Gabriel Migliarino, Matthias Orth, Sverre Sandberg, Marit S. Sylte, Sten Westgard and Elvar Theodorsson

The use of error and uncertainty methods in the medical laboratory

https://doi.org/10.1515/cclm-2017-0341
Received April 19, 2017; accepted July 4, 2017

Abstract: Error methods – compared with uncertainty methods – offer simpler, more intuitive and practical procedures for calculating measurement uncertainty and conducting quality assurance in laboratory medicine. However, uncertainty methods are preferred in other fields of science as reflected by the guide to the expression of uncertainty in measurement. When laboratory results are used for supporting medical diagnoses, the total uncertainty consists only partially of analytical variation. Biological variation, pre- and postanalytical variation all need to be included. Furthermore, all components of the measuring procedure need to be taken into account. Performance specifications for diagnostic tests should include the diagnostic uncertainty of the entire testing process. Uncertainty methods may be particularly useful for this purpose but have yet to show their strength in laboratory medicine. The purpose of this paper is to elucidate the pros and cons of error and uncertainty methods as groundwork for future consensus on their use in practical performance specifications. Error and uncertainty methods are complementary when evaluating measurement data.

Keywords: measurement uncertainty; performance specification; quality control; total error.

Introduction

The concept "total error" (TE) has different meanings to different authors and has also changed its definition over time [1, 2]. However, the widespread use of "total analytical error" (TAE) in laboratory medicine represent a testimony to its practical value.

The Milan conference on quality specifications resulted in a "consensus statement" published in 2015 [3], derived from the previous Stockholm consensus for performance specifications [4]. Recent developments contributing to the incentives for organizing the conference included the requirement of ISO 17025 and 15189 accreditation standards that laboratories routinely provide the measurement uncertainty (MU) of the results, the harmonization of the

Publishers Note: All authors are members of the Task and Finish Group on Total Error of the EFLM.

***Corresponding author: Wytze P. Oosterhuis,** Department of Clinical Chemistry and Haematology, Zuyderland Medical Center, Henri Dunantstraat 5, 6419 PC Heerlen, The Netherlands, Phone: +31 45 5766341, E-mail: w.oosterhuis@zuyderland.nl
Hassan Bayat: Sina Medical Laboratory, Qaemshahr, Iran. http://orcid.org/0000-0002-9958-9358
David Armbruster: Abbott Laboratories, Conway Park, Abbott Park, IL, USA
Abdurrahman Coskun: Acibadem University, School of Medicine, Department of Medical Biochemistry, Istanbul, Turkey
Kathleen P. Freeman: IDEXX Laboratories, Ltd, Grange House, Sandbeck Industrial Estate, Wetherby, West Yorkshire, UK
Anders Kallner: Department of Clinical Chemistry, Karolinska University Hospital Stockholm, Stockholm, Sweden
David Koch: Emory University School of Medicine, Grady Memorial Hospital in Atlanta, GA, USA
Finlay Mackenzie: University Hospitals Birmingham NHS Foundation Trust, Institute of Research and Development, Birmingham, UK
Gabriel Migliarino: Gmigliarino Consultants, Buenos Aires, Argentina
Matthias Orth: Vinzenz von Paul Kliniken gGmbH, Institut für Laboratoriumsmedizin, Stuttgart, Baden-Wurttemberg, Germany
Sverre Sandberg: Norwegian Quality Improvement of Primary Care Laboratories (Noklus), Institute of Global Health and Primary Health Care, University of Bergen and Laboratory of Clinical Biochemistry Haukeland University Hospital, Bergen, Norway
Marit S. Sylte: University of Bergen and Laboratory of Clinical Biochemistry Haukeland University Hospital, Bergen, Norway
Sten Westgard: Westgard QC, Madison, WI, USA
Elvar Theodorsson: Department of Clinical Chemistry and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

evaluation of proficiency testing procedures and recent challenges to the TE theory including the calculation of allowable total error (ATE) [5]. It was acknowledged during the Milan conference that many issues including TAE methods remain unresolved and needed further development. An EFLM Task and Finish Group on total error (TFG-TE) was established for that purpose.

TAE methods are firmly rooted in laboratory medicine, but a transition to the MU methods has taken place in other fields of metrology. TAE methods are commonly intertwined with quality assurance, analytical performance specifications and Six Sigma methods [2, 6–9], which are only partially included in this paper [9–12].

The aim of the present paper is to fulfill the task of the TFG-TE: to present a proposal on how to use the TE concept and how to possibly combine measures of bias and imprecision in performance specifications. The theoretical and practical underpinning of TAE and uncertainty methods are presented as groundwork for future consensus on their use in practical performance specifications.

This paper presents the consensus as reached within the EFLM task group but does however not represent a general EFLM consensus.

Materials and methods

The Delphi method [13], widely used method for achieving convergence of opinions of experts, was used. Two Delphi iterations were used. After an initial discussion phase of 8 months in which different views were presented, literature was exchanged and monographs were written, 37 statements were selected from the submitted texts and presented together with appropriate explanations. Eleven participants were subsequently invited to state their agreements or disagreements with the statements in writing. The chair of the TFG acted as moderator (WO). The TGF met to discuss the manuscript at the Warsaw EFLM-UEMS congress, in September 2016.

Results

It was agreed to use the terminology of Joint Committee on Guides in Metrology (JCGM) as expressed in the VIM [14] and GUM [15]. The terminology of VIM is notably not neutral regarding uncertainty models and does not include a definition of TAE.

Measurement error or simply error is a property of a single measurement – "measured quantity value minus a reference quantity value" [14]. Random measurement error is the component of measurement error that in replicate measurements varies in an unpredictable manner, whereas the systematic error remains constant – or varies in a predictable manner. The concept of error assumes that the difference between the measurement result and the "true value" or reference quantity value can be calculated.

Guide to the expression of uncertainty in measurement (GUM) and International Vocabulary of Metrology (VIM) do not use the concept of "true value". This is among the fundamental reasons why the concepts of TAE and MU seem incompatible. In the error perspective, measurement is seen as a process aimed at discovering quantity values, which have an independent existence (the "true values"). The better they are approximated, the less will be the error in the measurement result. MU – in contrast – conceives of no preexisting references or true values; ultimately, measurement results should be characterized in terms of the belief a subject attributes to them, expressed as the uncertainty of such results [16].

In 1974, Westgard et al. [1] introduced the concept of TAE in an effort to provide a quantitative measure for the acceptability of measurement method performance especially for proficiency testing. Reference laboratories estimate imprecision and bias separately by replicate measurements. In clinical laboratories, however, patient- and quality assurance samples are routinely assayed only once. TAE in these circumstances depends on the combined effect of the random and systematic errors of the method, which is compared with a defined allowable or permissible total error (pTAE, or total error allowable). TAE defines the maximum error for patient results that a single result can show with a certain probability, e.g. 95% or 99%. TAE thus estimates the limits of an interval around the true value where measured analytical results can be found with a defined probability. The TAE model further assumes that the difference between the patients' result and the true value can be estimated primarily from results from proficiency testing or from internal quality assurance. For many quality assurance applications, there is no need for separate performance goals for bias and imprecision.

Lately, efforts have been made to expand the TAE concept to the evaluation of results of patient samples, including all phases of the total testing process [2, 8, 9]. However, several additional sources of errors influence patient samples compared with control samples, e.g. preanalytical factors, patient factors (e.g. posture), matrix factors (calibrators or control samples that differ from the matrix in the patient samples), and interferences (e.g. in hemolytic, icteric or lipemic samples, drugs).

The concepts of bias and imprecision

Bias is the difference between the average of measurements made on the same sample and its reference [14, 17]. References are of two types: the reference defining the hypothetical error-free “true value” and the pragmatic reference or target value with an assigned value, e.g. quality control material procedures. The primary choice for estimating analytical bias is the use of a certified reference material (CRM) [18]. Optimal reference materials are not always available, especially when the measured quantity cannot not be unequivocally defined. Comparison with a reference method [14] can also be used for bias estimation. Sometimes, neither a CRM nor a reference method is available [19, 20], making a case for a reference that has a value assigned to it by agreement [20].

It is important to distinguish between short-term bias (e.g. within day, one shift) and long-term bias (e.g. during several weeks or months): many effects causing short-term bias, e.g. recalibrations may be seen as bias within this short time frame but may be indistinguishable from random effects when variation is observed over a longer period. When uncorrected, many short-term bias components increasingly contribute to the random error component of the MU.

In the case where samples of a particular patient are semirandomly allocated within a larger laboratory organization to different laboratories, several bias components will affect the result. This randomness may also be indistinguishable from imprecision (Figure 1).

Bias is included in the conventional TAE model as a constant without uncertainty. MU methods take this uncertainty into account. It is composed of the uncertainty of the values of the bias and of the reference standard. To improve the TAE estimation, the uncertainty of the bias was suggested to be incorporated into TAE calculation (note 3).

Precision is a concept of quality that is estimated quantitatively as its opposite – imprecision – and expressed as standard deviation or coefficient of variation (CV) [17]. Imprecision is estimated along a gradient between two extremes depending on the measuring conditions. One extreme is the repeatability condition where it is estimated by the same operator under conditions of the same laboratory, apparatus, method, material and within a short period. The other extreme is the reproducibility condition where the same sample is measured during extended periods (weeks, months or years) in different laboratories, involving different operators and measuring systems, methods, laboratory environments, management and quality assurance policies, and even different test methods. Intermediate conditions are conditions in between the two extremes, which need to be specified (see ISO 5725-1:1994).

Performance specifications

Westgard and Barry described performance specifications in terms of the TAE that can be tolerated in a test

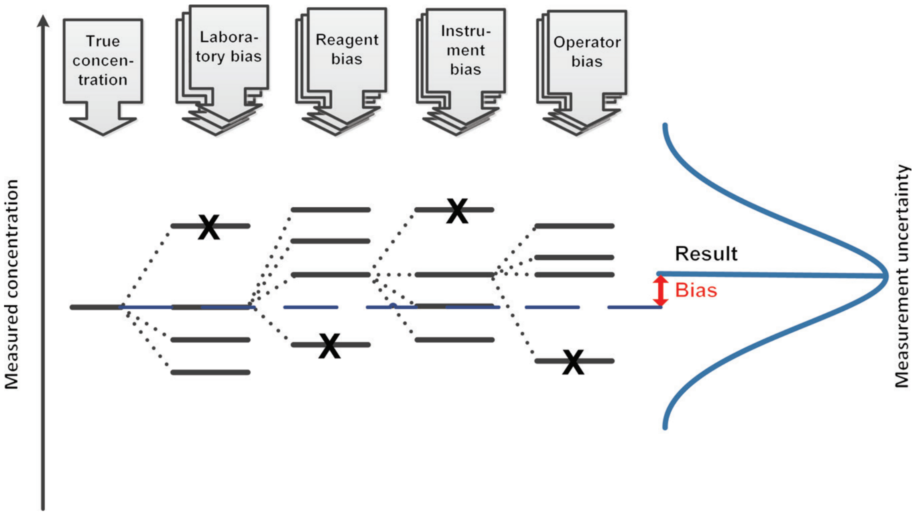

Figure 1: In the common situation where the samples of a particular patient are semirandomly allocated to different laboratories, reagents or reagent lots, measurement systems and operators, several bias components affecting the (hypothetical) true concentration value need to be considered and dealt with.
This randomness may be indistinguishable from imprecision.

result without compromising its medical usefulness [21]. Cembrowski and Carey [22] referred to quality goals in terms of the maximum clinically allowable error in a test result. When this paradigm is used to evaluate QC strategies, the primary performance measure of interest should be related to the probability of reporting a test result that contains an analytical error that exceeds the pTAE specification [23].

Recently, criteria were formulated to assign measurands to appropriate models for analytical performance specifications [24]. The preferred performance specification model is based on clinical needs. However, this model can at present only be applied for only a few measurands (e.g. HbA_{1c} and cholesterol). The second model is based on (components of) biological variation and can be applied for measurands that are in steady state or can be "transformed" to a steady-state situation in biological fluids [24]. A third model is based on the state of the art and can be applied in cases where models 1 and 2 cannot be used.

Since the study of Tonks [25], several alternative formulas have been suggested for calculating performance specifications based on biological variation. A review of a number of models based on between- and within-subject biological variation led to the notion, "A striking feature is the fact that all of the individual approaches recommend numbers for analytical standard deviation near or equal to 0.5 times the biological standard deviation" [26]. Expressed as analytical and biological coefficients of variation, $CV_A < 0.5CV_B$. In the case of monitoring, CV_I is used instead of CV_B.

Some measurands are subject to tight homeostatic control (e.g. electrolytes), lending themselves to very strict performance specifications. Three quality levels are used for imprecision and bias [27].

	Imprecision	Bias
Optimum quality	$CV_A \leq 0.25CV_I$	$\lvert B \rvert \leq 0.125CV_B$
Desirable quality	$CV_A \leq 0.5CV_I$	$\lvert B \rvert \leq 0.25CV_B$
Minimum quality	$CV_A \leq 0.75CV_I$	$\lvert B \rvert \leq 0.375CV_B$

These specifications do not combine specifications for imprecision and bias. The term for bias is equal to that derived from the model of Gowans et al. [28] (see below).

Performance specifications set limits for a test to establish whether the test is acceptable for routine use. Different specifications are needed for screening, diagnosis and monitoring. This also includes optimum goals that may be unachievable by current state-of-the-art procedures. It is vital to obtain a tool for defining the ideal specifications without the influence of the actual (or state of the art) analytical quality, as this will set a goal for manufacturers.

It might seem logical to apply the same limits for analytical performance as limits for internal quality assurance. However, there are reasons to set different limits: quality assurance applies rules to achieve a high probability of error detection and low probability of false rejection based on a singleton result. Quality assurance limits will generally be stricter than performance limits in order to maintain the performance goals and assure that – within a predefined probability – these goals are achieved.

Combining bias and imprecision and calculation of pTAE

Combining bias and imprecision specifications to set quality limits based on biological variation was proposed by Fraser and Petersen in 1993 [29, 30]:

$$\mathrm{pTAE} = 0.25(CV_I^2 + CV_G^2)^{1/2} + 1.65(0.5CV_I) \qquad (1)$$

The performance specification was proposed for proficiency testing but has been extensively used to define specifications for other purposes [31]. The term for bias is again that of Gowans et al. [28], combined with the generally accepted maximum imprecision of $0.5CV_I$ (with a coverage factor of 1.65, 95% one-sided). This expression defines a constant value for pTAE and shows a linear relationship between bias and imprecision.

In contrast with this conventional linear model is the curved model proposed by Gowans et al. [28]. With the transferability of reference intervals as starting point, the permissible bias and imprecision are calculated based on the premise that the reference interval limits will remain valid. Because of the inclusion of the biological variation in the model, the resulting relationship between maximum permissible bias and imprecision is curved (see Figure 2) where the related model of Larsen et al. [32] is used.

The model of Gowans and other reference limit-based models define reference limits based on biological variation alone as a simplification. Oosterhuis and Sandberg [33] adapted the model of Gowans et al. [28] to include the influence of analytical variation on the reference interval. Even inclusion of analytical variation, however, might lead to an underestimation of the actual reference interval limits, e.g. because of preanalytical variation (note 2). As an alternative, the actual reference interval limits can be used as starting point in the model [34]. It should be noted that in most distributions, CV_G and CV_I are log-Gaussian,

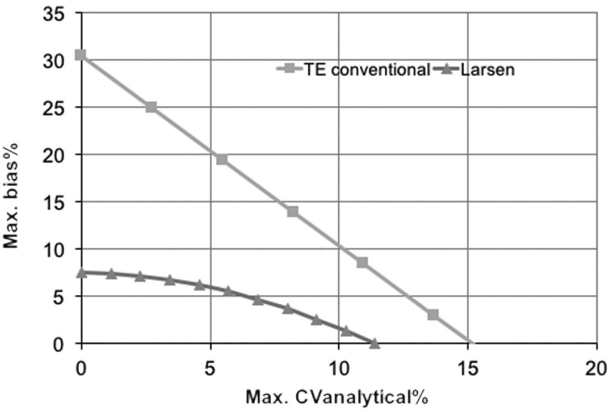

Figure 2: Relationship between maximum permissible imprecision and bias in the case of CK.
The conventional model ($z=2$, 2.3% outside limit, one sided) shows a linear relationship that leads to a constant value for pTAE (30%, equal to the bias at $CV_A=0$). In the model including biological variation, there is a curved relationship between imprecision and bias. pTAE is not a constant in this model (in this example: range 7.5% ($CV_A=0$) – 18.8% ($CV_A=11.4\%$) [32]. See note 4.

as are most reference ranges [35]. Performance specifications derived from biological data should ideally be based on this log-Gaussian distribution [36]. Other models combining bias and imprecision for calculation of pTAE might also be considered (e.g. [37–40]).

Considerations regarding the TE model

Two modes should be distinguished in proficiency testing and ICQ: the singleton and the multiple measurements of specimens. With a singleton IQC result, we only have the accuracy of the actual result: the difference of the measurement result and the target value (in other words the TAE). There is no way of distinguishing bias and random error components. This value has to be compared with a tolerance limit. However, during extended periods, the laboratory can in retrospect distinguish between bias and imprecision. This information can be derived from quality assurance results (both IQC and appropriately designed proficiency testing programs). In that perspective, bias and imprecision should be (and are) evaluated separately. Reduction of bias and imprecision call for different measures.

Essential statistical considerations govern the possibility to combine bias and imprecision. Bias has a sign, either positive or negative, whereas the standard deviation represents an interval of quantity values. In the CLSI-EP21, the absolute value of the bias is suggested to always be added to the imprecision [41]. In Rili-BAEK, the squared variance and a squared bias are added before taking the square root of the sum [12, 42]. A statistically correct addition of orthogonal variances using the Pythagorean theorem depends on the elimination of bias – as demanded by GUM.

The caveat with using a single statistic for the maximum permissible error is that the impact or acceptability of its different components is unknown, although bias and imprecision work out quite differently. The same TAE value can result from a low bias and a high imprecision and vice versa. Based on models for diagnosis (e.g. HbA_{1c} in diabetes), bias and imprecision have quite different effects on the proportion of false positives and false negatives [42]. HbA_{1c} is a good example for which there are separate goals for bias (±2%) and imprecision (3%). The pTAE of ±6% is used for practical purposes, e.g. as by the College of American Pathologist's for grading HbA_{1c} proficiency testing surveys. This data compression might not serve the purpose of goal setting or clinical performance description [12].

pTAE is frequently calculated based on biological variation [28, 43]. The conventional linear model (expression 1) has been criticized for a number of reasons [5]:

1. Both maxima of permissible bias and imprecision are added to obtain pTAE, a pragmatic solution first proposed for the use in proficiency testing [33]. The theoretical basis for this is lacking. Two maximum permissible errors are added, derived under the mutually exclusive assumptions of zero bias and zero imprecision, respectively. The sum will allow an increase of the percentage of test results exceeding the predefined limits.
2. The performance specification for imprecision is in general $CV_A<0.5CV_B$. In the case of diagnosis, this can be written as $CV_A<0.5(CV_I^2+CV_G^2)^{1/2}$. In case of monitoring, only the within-subject variation CV_I is included. The maximum permissible bias was derived as $0.25CV_B$ or $0.25(CV_I^2+CV_G^2)^{1/2}$. It should be noted, however, that in the conventional model, this bias term is applied in the case of monitoring although this expression had been derived from a reference range model and only applies to diagnosis. As an alternative, a model based on a reference change value model was developed that is only based on CV_I and not on CV_G [32].
3. The conventional linear TAE model takes into account only the analytical variation and the bias. When a tolerance limit is defined at zSD – assuming a Gaussian distribution – with a fixed number of test results outside this limit (e.g. 5% at $z=1.64$), the combinations of bias and imprecision (CV_A) fulfilling this condition

show a straight line with a slope of –z (see also Figure 2, TAE conventional with z = 2). At one extreme of the line, we have bias = 0, with CVa = pTAE/z; at the other extreme, we have $CV_A = 0$ (a hypothetical value) and bias = pTAE. The straight line has the characteristic that the tolerance limit pTAE, a combination of bias and imprecision, is a constant equal to the bias at $CV_A = 0$. This linearity is of fundamental importance in the methods advocated by both J. Westgard and S. Westgard (www.westgard.com). This model is only valid when imprecision and bias are the only variables involved, or in other words, when the distribution of test results is Gaussian and completely defined by the analytical bias and imprecision alone. Thus – in its original form – the TE model does not include biological variation and other additional causes of variation into the model, except in the calculation of pTAE [1]. Biological variation, although not relevant when monitoring variation in control samples, is certainly relevant when dealing with patient samples [10].

Biological variation can be taken into account by including biological variation in the model transforming the relationship between imprecision and bias into a curve (see Figure 2 Larsen et al. [32]). The consequence is that the value of pTAE is not a constant in this model for any combination of bias and imprecision (Figure 2).

4. It has been argued that the condition $CV_A < 0.5CV_I$ relates to performance specifications that cannot be maintained by internal quality assurance (e.g. leading to a sigma metric below 3, see note 1). For that reason, the 0.5 and the 0.25 coefficients for imprecision and bias might be questioned.

Measurement uncertainty

Uncertainty methods as endorsed by the BIPM [15] originated in physical measurements [44, 45], and chemistry was included as late as in the 1980s. Laboratories of chemistry and related sciences have struggled when adapting to a long tradition established by physical metrology laboratories [45, 46].

The basic parameter of MU is the SD, and the symbol for uncertainty is u. In practice, bias correction can reduce systematic errors, and replicate measurements can diminish the effect of random errors. However, this cannot completely eliminate these errors. For that reason – according to the MU concept – the "true value" cannot be exactly known. A measurement result represents the "best estimate" of the measured quantity. The combined uncertainties of bias and imprecision provide an interval of values within which the (unknowable, hypothetical) "true value" of the measured quantity is believed to lie, with a stated coverage probability (e.g. 95%). The MU concept also assumes that if the bias of a procedure is known, then steps are to be taken to minimize it, e.g. by recalibration. However, because the bias value cannot be known exactly, an uncertainty will be associated with such a correction. Thus, in the MU concept, a measurement result can comprise two uncertainties: the uncertainty associated with bias correction and the uncertainty due to imprecision. The uncertainties that act on the measurement result are combined to one MU statistic.

Current developments within the BIPM regarding future versions of GUM and VIM

According to GUM [15], MU reflects the lack of exact knowledge of the value of the measurand. Developments in MU also emphasize the relationship between the measurement itself and the theoretical models and philosophies underpinning the use of the measurement results [47–55].

The 1993/1995 version of the GUM catered for both uncertainty and error approaches [56]. Recent supplements to the GUM and the latest version of VIM (VIM3) published in 2008 went completely in the direction of the uncertainty approach, including Bayesian statistics that ultimately relate to the uncertainty of diagnosis [57–59]. However, inconsistencies between the supplements and the main GUM text have been pointed out repeatedly [60–62], shedding a light on an obvious need to harmonize the approaches to error and uncertainty. Professor Luca Mari, a long-time member of the JCGM Working Group on the International Vocabulary of Metrology (VIM), has in multiple recent publications, e.g. in [16, 47], made a case that error and uncertainty methods are not only compatible but also complementary when evaluating measurement data. Our working group finds this approach highly appropriate for use in medical laboratories.

It should be noted that models for the calculation of permissible maximum bias and maximum imprecision are not restricted to either MU or TAE, but the results of these calculations are applicable in both paradigms. In IQC, the TAE and MU models approach each other and can become similar, and the goal setting developed in the TAE model combining bias and imprecision is also applicable for MU [11].

According to ISO 15189 [63], MU should be made available by the laboratory on request. Careful interpretation is needed, as MU is already taken into account also in other test characteristics such as sensitivity, specificity, predictive value, etc. Physicians might correct decision levels knowing the MU of a test, taking MU into account again. Analytical variation as well as other sources of uncertainty, including biological and preanalytical variation, are already reflected in reference interval limits and are (or should be) also taken into account in decision limits in clinical guidelines, as "grey zones", and borderline areas. Lower MU results in higher predictive values and lower clinical uncertainty. This might even lead to a different clinical application of the test, with HbA_{1c} assays as a typical example. The improved reliability of the results made expert organizations revise the guidelines and recommend HbA_{1c} for diagnosis. Such an improved clinical uncertainty is also seen with high-sensitive troponin assays.

The problem of the unknown true value in relation to the definition of error is circumvented in VIM by defining the error with respect to the reference quantity value. This serves a "surrogate true value" within the system (or model), e.g. in the case of internal quality assurance or proficiency testing samples. In the case of patient test results, only an estimated value (or probability) can be assigned to the TE.

We attempt to understand the real world by modelling it. Within a model, we can measure values and make our interpretations within the assumptions of the model. Within the model, we do have "true" values. However, between the modeled values and the real world, there are many uncertainties.

Transposed to our laboratory, this translates to the following (Figure 3): our model consists of calibrators with assigned values and internal quality assurance materials with target values. These are our "true" values, and we can very well express error (trueness) as a number. Proficiency testing, reference methods and certified reference standards represent another model of another level.

However, the measurement of patient samples can be seen as an entity in the real world and knowing the "real" value of the measurands is impossible. Here we can follow the VIM approach, starting with a reference quantity value for our standards and calibrators. The reference quantity value (or reference standard) is determined by the agreed-upon reference method. Therefore, the "analytically" real value of a measurand in the patient sample is the value ultimately traceable to the reference method. The measurement procedure and other sources of variation will contribute to the combined MU of the final measurement result.

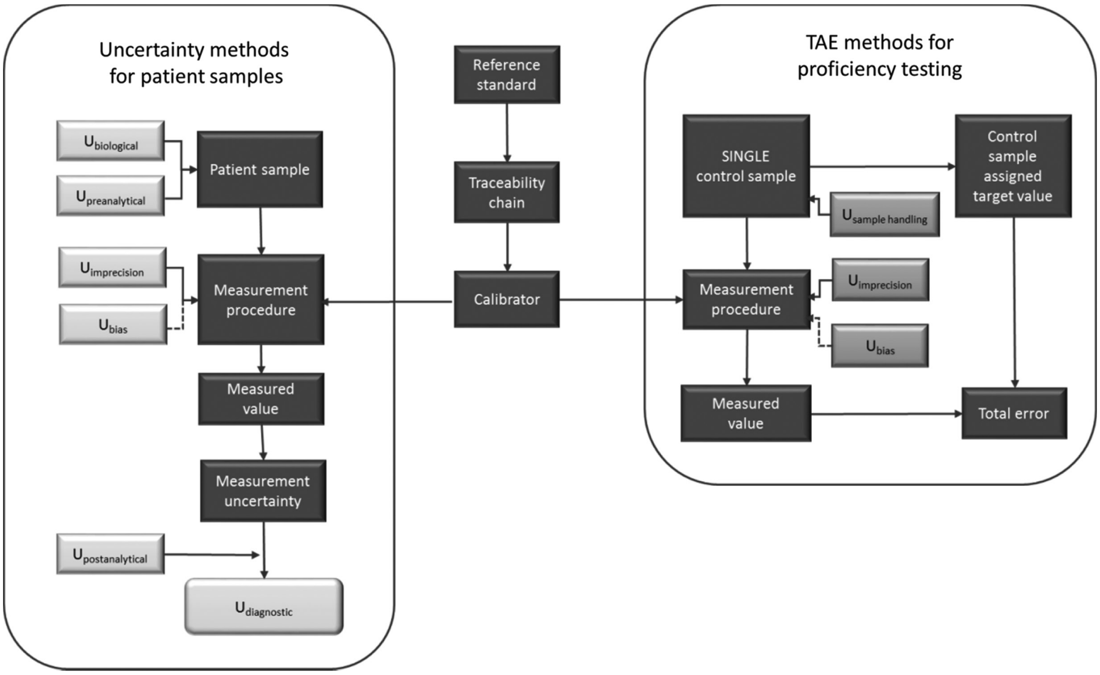

Figure 3: Illustration of the use of MU methods for patient samples and TAE methods for proficiency testing.
A target value is defined for the proficiency testing sample, which is used for calculating error. In patient samples, uncertainty methods estimate the confidence we can have in the measurement result for the purpose of diagnosis. Proficiency testing and measurement uncertainty are related through the traceability chain to the reference standard. The dotted lines indicate efforts to eliminate bias.

Both TAE and MU theory are related to traceability. “Comparability of results across laboratories” must be the goal of the TAE concept. The bias concept in the TAE is in fact related to “traceability”: TAE relates to the difference of the results with a true (reference) value and in this way includes the traceability of the results. What is ignored in the traditional TAE is the degree of the reliability of the transference of the assigned value through the traceability chain and also the reliability of the experiment used for the determination of bias. MU, through incorporating these sources of uncertainty, can help TAE to be an improved measure of traceability of results.

Discussion

TAE methods for quantifying the quality of measuring systems and defining performance specifications are widely used in laboratory medicine. They are appropriate for analyzing data from single measurement results in proficiency testing schemes and may constitute a basis for the calculation of performance.

The ultimate aim of measuring patient samples in laboratory medicine is to improve the understanding of possible disease conditions or to monitor treatment effects. This aim is influenced by the uncertainties not only of the measuring system but also by biological, preanalytical and postanalytical uncertainty components. This is represented in the Bayesian model used by MU that takes into account all uncertainty components, including bias and imprecision, that effect the measurement result.

Although TAE is generally understood as total analytical error, the types of errors included in the TE definition are varied: preanalytical errors, biological variation and postanalytical errors. It has been argued that the TE concept only deserves the predicate “total” when all kinds of errors are included [11]. Notably, the permissible TAE is often derived from reference intervals [36, 64, 65] that are substantially influenced by other types of errors (variation) than the analytical variation.

Recent developments in the philosophy of measurement sciences, including those within the BIPM/JCGM, indicate that more than one philosophical outlook and a corresponding selection of different measurement uncertainty calculations may be endorsed. Therefore, it seems prudent to continue the use of TAE methods and when appropriate replace them with MU calculations when the latter offer proven advantages.

Among the major challenges that error methods face are the following:

1. There are several variants in how the calculations of TAE are being performed and used, with flaws in pTAE calculations [5].
2. The concept of the “true value” has been abandoned in metrology. If a true value cannot be known, TAE cannot be estimated. We must use “surrogate true values” to estimate TAE.
3. TE models do not appreciate the uncertainty of bias estimation/correction.

MU methods in laboratory medicine also face several challenges:

1. Error methods are well understood and widely implemented in laboratories. Procedures for implementing them in ISO accreditation schemes are well accepted by accreditation authorities. Therefore, there are no incentives for leaving error methods in favor of uncertainty methods.
2. Methods to calculate “permissible MU” as well as quality assurance procedures based on MU theory are not well developed.
3. Unfamiliarity with uncertainty calculations within most laboratories [66].
4. The level of knowledge regarding other causes of variation than analytical variation including biological variation, preanalytical and postanalytical variation needed for the estimation of MU is still limited.

Uncertainty methods, according to GUM, are in the early phases of implementation in laboratory medicine in Australia [40, 66, 67]. The majority of other countries rely on classic top-down uncertainty calculations, sometimes using ANOVA, covariance analysis and variance component analysis [68–70].

Imprecision can be estimated as repeatability imprecision at the one extreme, and reproducibility imprecision at the other, with intermediate imprecisions in between. Similarly, bias needs to be estimated separately in the context of the time frame and also the complexity of the measuring systems, including several measuring systems in different laboratories within a laboratory organization (Figure 1). Quality assurance in laboratory medicine needs to be more comprehensively developed to address a situation far more complicated than a single analyzer working in batch mode, for which it was originally developed. The current situation includes large laboratory conglomerates with multiple analyzers working continuously. The most important tasks for laboratory conglomerates are (1) to reduce between-analyzer bias and (2) to identify and technically correct measuring systems with excessive imprecision. Performance specifications should be related to the properties of a whole diagnostic organization including all its measuring systems for diagnosing and monitoring patients. These performance specifications are not yet sufficiently developed despite the fact that ISO 17025 and 15189 require laboratories and conglomerates of laboratories to estimate the overall uncertainty.

Bias and imprecision are different error types with different causes that obviously require different means of correction. Medical laboratories should therefore establish and maintain routines for estimating and minimizing them separately. Bias, however, remains a complicating factor. MU methods advocate correcting bias and include the uncertainty of bias correction. Short-term bias may be indistinguishable from random effects when variation is observed over extended periods and may contribute to the random error component of the MU. There are many methods for combining bias and imprecision, but each method represents some kind of concession, assumption or reduction. Further research and development in this area is needed in order to establish consensus on methods that are optimal for medical laboratories.

Analytical performance specifications should take the diagnostic uncertainty of the whole testing process into account. MU methods according to GUM have not been sufficiently developed to deal with diagnostic uncertainty in laboratory medicine. Development of analytical performance specifications for diagnostic uncertainty has the potential of creating a paradigm shift in laboratory medicine resulting in quality improvements and improved use of diagnostic methods. In anticipation of this paradigm shift, error methods remain the most used methods for quality assurance and analytical performance specifications in laboratory medicine. Currently, error and uncertainty methods are complementary when evaluating measurement results in medical laboratories.

Acknowledgments: Bossuyt P. M., Deman P., Faria A. P., Ritzen H., Surtihadi J. and Hyltoft Petersen P.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: 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.

Notes

Note 1. Calculation of sigma metric.

$$\begin{aligned}
\text{ATE} &= \text{Bias} + 1.65(0.5\text{CV}_B)\\
\text{Sigma metric} &= (\text{ATE-Bias})/\text{CV}_A\\
&= (\text{Bias} + 1.65(0.5\text{CV}_B))/\text{CV}_A \quad \text{with bias} = 0\\
&= (0.25\text{CV}_B + 0.825\text{CV}_B)/\text{CV}_A\\
&= 1.075\text{CV}_B/\text{CV}_A \quad \text{with } \text{CV}_A = 0.5\text{CV}_B\\
&= 2.15\ \text{CV}_A/\text{CV}_A\\
\text{Sigma metric} &= 2.15
\end{aligned}$$

Note 2. An example is sodium, with: $\text{CV}_I = 0.6\%$; $\text{CV}_G = 0.7\%$, $\text{CV}_A = 0.58\%$, the calculated reference interval (with $\text{CV}_{tot} = \ \text{CV}_I^2 + \text{CV}_G^2 + \text{CV}_A^2$) is 137–143 mmol/L. The actual reference interval of this laboratory is 135–145 mmol/L, a 40% difference that will be reflected in the pTAE. The extra variation will be caused by preanalytical and other factors.

Note 3. Here is where the term uncertainty of bias (U_B) needs to be addressed. The difference between the traditional TAE and MU is that MU adds U_B to the SD. We can use U_B in the TAE calculation to get an estimation of TE as (U_c = uncertainty of concentration):

$$\text{TAE} = \text{B} + \text{k} * (U_B^2 + \text{SD}_A^2)^{0.5}$$

$$\text{TAE} = \text{B} + \text{k} * U_c$$

$$\text{TAE} = \text{B} + \text{MU}$$

If bias is zero (either absent from the beginning or removed via recalibration or correction), then the previous equation becomes TAE = MU. According to MU concept, bias should be corrected when known, and the uncertainty of this correction was included.

Note 4. Strictly speaking, these models are difficult to compare: The model of Larsen is aimed at monitoring and only includes CV_I. The other model – although used for monitoring – includes both CV_I and CV_G as discussed in section, which is considered a flaw in this model.

References

1. Westgard JO, Carey RN, Wold S. Criteria for judging precision and accuracy in method development and evaluation. Clin Chem 1974;20:825–33.
2. Westgard JO. Useful measures and models for analytical quality management in medical laboratories. Clin Chem Lab Med 2016;54:223–33.

3. 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.
4. Kenny D, Fraser CG, Petersen PH, Kallner A. Consensus agreement. Scand J Clin Lab Inv 1999;59:585.
5. Oosterhuis WP. Gross overestimation of total allowable error based on biological variation. Clin Chem 2011;57:1334–6.
6. Westgard JO, Westgard SA. Assessing quality on the Sigma scale from proficiency testing and external quality assessment surveys. Clin Chem Lab Med 2015;53:1531–5.
7. Westgard SA. Utilizing global data to estimate analytical performance on the Sigma scale: a global comparative analysis of methods, instruments, and manufacturers through external quality assurance and proficiency testing programs. Clin Biochem 2016;49:699–707.
8. Westgard S. Quality Goals at the Crossroads: Growing, Going, or Gone? 2016. Available at: http://www.westgard.com/gone-goals-gone.htm. Accessed: 30 July 2017.
9. Westgard JO, Westgard SA. Quality control review: implementing a scientifically based quality control system. Ann Clin Biochem 2016;53(Pt 1):32–50.
10. Panteghini M, Sandberg S. Total error vs. measurement uncertainty: the match continues. Clin Chem Lab Med 2016;54:195–6.
11. Oosterhuis WP, Theodorsson E. Total error vs. measurement uncertainty: revolution or evolution? Clin Chem Lab Med 2016;54:235–9.
12. Kallner A. Is the combination of trueness and precision in one expression meaningful? On the use of total error and uncertainty in clinical chemistry. Clin Chem Lab Med 2016;54:1291–7.
13. Dalkey NC, Helmer O. An experimental application of the Delphi method to the use of experts. Manag Sci 1963;9:458–67.
14. JCGM. International vocabulary of metrology – basic and general concepts and associated terms (VIM 3): Bureau International des Poids et Mesures; 2012. [3 edition] Available at: http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf. Accessed: 30 July 2017.
15. 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. Accessed: 30 July 2017. Joint Committee for Guides in Metrology, 2008.
16. Mari L. The 'error approach' and the 'uncertainty approach': are they incompatible? Leiden: Lorentz Center, 2011.
17. Menditto A, Patriarca M, Magnusson B. Understanding the meaning of accuracy, trueness and precision. Accred Qual Assur 2007;12:45–7.
18. De Bievre P. On 'trueness control materials', better known under the multi-purpose term of 'Certified Reference Materials' (CRMs). Accredit Qual Assur 2010;15:71–2.
19. Vesper HW, Miller WG, Myers GL. Reference materials and commutability. Clin Biochem Rev 2007;28:139–47.
20. Greg Miller W, Myers GL, Lou Gantzer M, Kahn SE, Schonbrunner ER, Thienpont LM, et al. Roadmap for harmonization of clinical laboratory measurement procedures. Clin Chem 2011;57:1108–17.
21. Westgard JO, Seehafer JJ, Barry PL. Allowable imprecision for laboratory tests based on clinical and analytical test outcome criteria. Clin Chem 1994;40:1909–14.
22. Cembrowski GS, Carey RN. Considerations for the implementation of clinically derived quality control procedures. Lab Med 1989;20:400–5.
23. Parvin CA. Quality-control (QC) performance measures and the QC planning process. Clin Chem 1997;43:602–7.
24. Ceriotti F, Fernandez-Calle P, Klee GG, Nordin G, Sandberg S, Streichert T, et al. Criteria for assigning laboratory measurands to models for analytical performance specifications defined in the 1st EFLM Strategic Conference. Clin Chem Lab Med 2017;55:189–94.
25. Tonks DB. Quality control systems in clinical chemistry laboratories. Postgrad Med 1963;34:A58–70.
26. Stöckl D, Baadenhuijsen H, Fraser CG, Libeer J-C, Petersen PH, Ricós C. Desirable routine analytical goals for quantities assayed in serum. Discussion paper from the members of the external quality assessment (EQA) working group A on analytical goals in laboratory medicine. Eur J Clin Chem Clin Biochem 1995;33:157–69.
27. Fraser CG, Petersen PH, Libeer JC, Ricos C. Proposals for setting generally applicable quality goals solely based on biology. Ann Clin Biochem 1997;34(Pt 1):8–12.
28. Gowans EM, Peteresen PH, Blaabjerg O, Hörder M. Analytical goals for the acceptance of common reference intervals for laboratories throughout a geographical area. Scand J Clin Lab Invest 1988;48:757–64.
29. Fraser CG, Petersen PH. Quality goals in external quality assessment are best based on biology. Scand J Clin Lab Invest 1993;53:8–9.
30. Fraser CG, Peterson PH. Desirable standards for laboratory tests if they are to fulfill medical needs. Clin Chem 1993;39:1447–5.
31. Ricos C, Alvarez V, Cava F, Garcia-Lario JV, Hernandez A, Jimenez CV, et al. Desirable specifications for total error, imprecision, and bias, derived from intra- and inter-individual biologic variation. Available at: http://www.westgard.com/biodatabase1.htm. Accessed: 30 July 2017.
32. Larsen ML, Fraser CG, Petersen PH. A comparison of analytical goals for haemoglobin A1c assays derived using different strategies. Ann Clin Biochem 1991;28(Pt 3):272–8.
33. Oosterhuis WP, Sandberg S. Proposal for the modification of the conventional model for establishing performance specifications. Clin Chem Lab Med 2015;53:925–37.
34. Haeckel R, Wosniok W, Postma T. Quantity quotient reporting. Comparison of various models. Clin Chem Lab Med 2015;53:1921–6.
35. Petersen PH, Stockl D, Blaabjerg O, Pedersen B, Birkemose E, Thienpont L, et al. Graphical interpretation of analytical data from comparison of a field method with reference method by use of difference plots. Clin Chem 1997;43:2039–46.
36. Haeckel R, Wosniok W, Gurr E, Peil B. Permissible limits for uncertainty of measurement in laboratory medicine. Clin Chem Lab Med 2015;53:1161–71.
37. Asberg A, Odsaeter IH, Carlsen SM, Mikkelsen G. Using the likelihood ratio to evaluate allowable total error – an example with glycated hemoglobin (HbA(1c)). Clin Chem Lab Med 2015;53:1459–64.
38. Oddoze C, Lombard E, Portugal H. Stability study of 81 analytes in human whole blood, in serum and in plasma. Clin Biochem 2012;45:464–9.

39. McDonald R. Quality assessment of quantitative analytical results in laboratory medicine by root mean square of measurement deviation. J Lab Med 2006;30:111–7.
40. White GH, Farrance I, AACB Uncertainty of Measurement Working Group. Uncertainty of measurement in quantitative medical testing: a laboratory implementation guide. Clin Biochem Rev 2004;25:S1–24.
41. Astles JR, Person NB, Armbruster DA, Pierson-Perry JF, Kondratovich MV, Scott MG, et al. CLSI-EP21, Evaluation of total analytical error for quantitative medical laboratory measurement procedures, 2nd ed. Wayne, NJ: Clinical and Laboratory Standards Institute, 2016.
42. Petersen PH, Klee GG. Influence of analytical bias and imprecision on the number of false positive results using guideline-driven medical decision limits (Reprinted from Clinica Chimica Acta vol 430C, pg 1–8, 2014). Clin Chim Acta 2014;432:127–34.
43. Fraser CG, Petersen PH. Desirable standards for laboratory tests if they are to fulfill medical needs. Clin Chem 1993;39:1447–53; discussion 53–5.
44. Page CH, Vigoureux PE. The International Bureau of Weights and Measures 1875–1975. US Department of Commerce, National Bureau of Standards, Gaithersburg, USA, NBS Special Publication 420, May 1975, Available at: https://archive.org/details/internationalbur 420page. Accessed: 27 July 2017.
45. Williams A. What can we learn from traceability in physical measurements? Accredit Qual Assur 2000;5:414–7.
46. Williams A. Traceability and uncertainty – a comparison of their application in chemical and physical measurement. Accredit Qual Assur 2001;6:73–5.
47. Mari L, Giordani A. Modeling measurement: error and uncertainty. In: Boumans M, Hon G, Petersen A, editors. Error and uncertainty in scientific practice. London: Pickering & Chatto, 2014:79–96.
48. Giordani A, Mari L. Measurement, models, and uncertainty. Ieee T Instrum Meas 2012;61:2144–52.
49. Psillos S. Scientific realism: how science tracks truth. London: Routledge, 1999.
50. Giere RN. Explaining science: a cognitive approach. Chicago: University of Chicago Press, 1988.
51. Giere RN. Cognitive models of science. Minneapolis: University of Minnesota Press, 1992:xxviii:508.
52. Giere RN. Scientific perspectivism. Chicago: University of Chicago Press, 2006:151.
53. Tal E. Old and new problems in philosophy of measurements. Philosophy Compass 2013;8:1159–73.
54. Tal E. Measurement in Science. 2015. In: The Stanford Encyclopedia of Philosophy [Internet]. Available at: http://plato.stanford.edu/archives/sum2015/entries/measurement-science. Accessed: 30 July 2017.
55. Tal E. Measurement in science. In: Zalta EN, editor. Stanford encyclopedia of philosophy, 2015.
56. Kacker R, Jones AW. On use of Bayesian statistics to make the guide to the expression of uncertainty in measurement consistent. Metrologia 2003;40:235–48.
57. Vallverdú J. Bayesian versus frequentists. A philosophical debate on statistical reasoning. Heidelberg: Springer, 2016.
58. Sanogo M, Abatih E, Saegerman C. Bayesian versus frequentist methods for estimating true prevalence of disease and diagnostic test performance. Vet J 2014;202:204–7.
59. Weise K, Woger W. A Bayesian theory of measurement uncertainty. Meas Sci Technol 1993;4:1–11.
60. Lira I. The GUM revision: the Bayesian view toward the expression of measurement uncertainty. Eur J Physics 2016;37:1–16.
61. Kyriazis GA. Contributions to the revision of the 'Guide to the expression of uncertainty in measurement'. J Phys Conf Ser 2015;575:1.
62. Bich W. How to revise the GUM? Accredit Qual Assur 2008;13:271–5.
63. ISO. ISO 15189:2012 Medical laboratories – requirements for quality and competence. Geneva: International Standardisation Organisation, 2012.
64. Haeckel R, Wosniok W, Streichert T. Optimizing the use of the "state-of-the-art" performance criteria. Clin Chem Lab Med 2015;53:887–91.
65. Haeckel R, Wosniok W. A new concept to derive permissible limits for analytical imprecision and bias considering diagnostic requirements and technical state-of-the-art. Clin Chem Lab Med 2011;49:623–35.
66. Farrance I, Frenkel R. Uncertainty of measurement: a review of the rules for calculating uncertainty components through functional relationships. Clin Biochem Rev 2012;33:49–75.
67. Farrance I, Frenkel R. Uncertainty in measurement: a review of monte carlo simulation using microsoft excel for the calculation of uncertainties through functional relationships, including uncertainties in empirically derived constants. Clin Biochem Rev 2014;35:37–61.
68. Norheim S. Computer support simplifying uncertainty estimation using patient samples. Linkoping: Linkoping University, 2008. Available at: http://liu.diva-portal.org/smash/record.jsf?pid=diva2:417298. Accessed: 30 July 2017.
69. Theodorsson E. Validation and verification of measurement methods in clinical chemistry. Bioanalysis 2012;4:305–20.
70. Theodorsson E, Magnusson B, Leito I. Bias in clinical chemistry. Bioanalysis 2014;6:2855–75.