Clinica Chimica Acta 495 (2019) 129–138

Contents lists available at ScienceDirect

Clinica Chimica Acta

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

Review

Bias in analytical chemistry: A review of selected procedures for incorporating uncorrected bias into the expanded uncertainty of analytical measurements and a graphical method for evaluating the concordance of reference and test procedures

Robert Frenkel[a,*,1], Ian Farrance[b], Tony Badrick[c]

[a] 96 Shirley Road, Roseville, New South Wales 2069, Australia
[b] Discipline of Laboratory Medicine, School of Health and Biomedical Sciences, RMIT University, Bundoora, Victoria 3083, Australia
[c] RCPA Quality Assurance Programs, Suite 201, 8 Herbert Street, St Leonards, NSW 2065, Australia

ARTICLE INFO

Keywords:
Measurement uncertainty
Bias
Bias uncertainty
Bias correction
Probability density function
Overlap

ABSTRACT

The *Evaluation of measurement data - Guide to the Expression of Uncertainty in Measurement* (GUM) provides the framework for evaluating measurement uncertainty. The preferred GUM approach for addressing bias assumes that all systematic errors are identified and corrected at an early stage in the measurement process. We review some procedures for treating uncorrected bias and its inclusion into an overall uncertainty statement. When bias and its uncertainty are recognised as metrological states independent of scatter in the test results, the uncertainty of the reference and uncertainty of the bias can be equated. The net standard uncertainty of a test result is the root-sum-square of the standard uncertainty of the bias and the standard uncertainty of measurements on the test. Since an incomplete and therefore potentially erroneous formula is often used for estimating bias standard uncertainty, we propose an alternative calculation. We next propose a graphical method using a simple algorithm that quantifies the discrepancy between the results of a test measurement and the corresponding reference value, in terms of the percentage overlap of two probability density functions. We propose that bias should be corrected wherever possible and we illustrate this approach using the graphical method. Even though this review is focused principally on analytical chemistry and medical laboratory applications, much of the discussion is applicable to all areas of metrology.

1. Introduction

Measurement error comes in two forms, *systematic* and *random*. A systematic measurement error, or *bias*, is the component of total measurement error that in 'replicate measurements (under the same environmental conditions) remains constant or varies in a predictable manner' [1]. Since its introduction in 1993, the *Evaluation of measurement data – Guide to the Expression of Uncertainty in Measurement'* (GUM) [2] has provided the basic framework for evaluating measurement uncertainty (MU). The preferred GUM approach for addressing the issue of bias is based on the assumption that all systematic errors are identified and corrected at an early stage in the measurement process. The accuracy of a measurement then depends on the uncertainty associated with random errors combined with the uncertainty associated with any correction for systematic error (bias). However, section F2.4.5 of the GUM specifically describes the procedure where a significant uncorrected bias may be taken into account by enlarging the uncertainty assigned to the result.

As a matter of terminology and assumptions in this paper, we note the following six points.

1. The term *standard uncertainty*, denoted in general by a lower-case u, is synonymous with both *standard deviation* and *imprecision*. The square of this quantity, u^2, is the *variance*. When a standard uncertainty is calculated from n replicate measurements, the standard uncertainty of the *mean* of these measurements is often of interest (as in Sections 2.1, 2.5 and 3). Provided that the replicate measurements are mutually uncorrelated, the standard uncertainty of their mean is $u/\sqrt{n}$.
2. Probability density functions (PDFs) play a major role in this paper.

[*] Corresponding author.
E-mail addresses: frenkelfamily@hotmail.com (R. Frenkel), ian.farrance@rmit.edu.au (I. Farrance), Tony.Badrick@rcpaqap.com.au (T. Badrick).
[1] formerly of the National Measurement Institute Australia, Bradfield Road, West Lindfield, New South Wales 2070, Australia.

https://doi.org/10.1016/j.cca.2019.03.1633
Received 20 February 2019; Received in revised form 26 March 2019; Accepted 27 March 2019
Available online 29 March 2019

In any PDF, the quantity x that is being measured is plotted along the horizontal axis. For each value of x, there is a corresponding probability density $y = f(x)$. The quantity y is the height of the curve vertically above x. Essentially, the curve shows how the variable x is distributed over its range of values and shows where x is most likely (large y) and least likely (small y) to be located. It is emphasised that the quantity y on the vertical axis is not probability, but probability density. The probability that x should be located within given limits is the *area* under the curve between those limits. The relationship between probability density and probability is analogous to the relationship between ordinary physical density and mass. It is also important to note that since x must have *some* value within its permitted range, the total area under the curve of a PDF must be 1, or certainty.

3. We assume, unless otherwise stated, that variables such as the estimated values of measurand and of bias follow a Gaussian PDF. More details about a Gaussian PDF are given in the Supplementary Appendix.
4. For Gaussian PDFs, where bias is insignificant or has been corrected, *expanded uncertainty*, denoted in general by an upper-case U, is the standard uncertainty u multiplied by a *coverage factor*, k, which for the usual figure of a 95% confidence range is $k = 1.96$. In what follows we make the approximation $k \approx 2$, except where otherwise stated. The confidence range $\pm U = \pm ku = \pm 2u$ is symmetric about the estimated value of the measurand, and this estimated value is located within the range $\pm U$ with 95% probability. Where bias has not been corrected, the relationship between U and u is more complicated and may involve the value of bias itself. Nevertheless, U continues to be interpreted as a confidence range, with a target figure of 95% for the estimated value of the measurand, and will generally be symmetric about that estimated value.
5. Quantities expressed in relative terms, for example as percentages, will be denoted with '(rel)' after the quantity. Where a relative value of standard uncertainty u(rel) occurs as a squared term, namely as a variance, this will be written numerically without any percentage sign.
6. The term *reference* refers to a primary source such as a certified reference material (CRM), a standard reference material (SRM), a reference method or reference procedure used as the reference in a method comparison study. The reference is the standard for comparison measurements using a *test* system, the results of which will be compared with the reference. If the mean of the test results is significantly different from the reference, a systematic error or bias exists. The correction for bias, if carried out, will then involve adjusting the test result (or mean test result) to match that of the reference.

As we see the situation, there are two points of view involved in the observance of the GUM: first, the pure metrological and statistical problem of incorporating uncorrected bias into an overall uncertainty statement; and, second, the practical aspects of expanding the range of uncertainty within which a particular result may be located. Whereas the first point of view requires strict statistical procedures to be observed, in the second more empirical point of view, an expanded uncertainty is estimated using practically feasible and agreed procedures. While GUM prefers all bias to be eliminated or corrected, the practical aspects of applied analytical chemistry allow an expanded uncertainty range which incorporates both systematic and random uncertainty components. The contentious nature of establishing such an expanded uncertainty range is demonstrated by the numerous important and helpful articles [3–24] describing bias and proposed procedures for its inclusion into the overall uncertainty statement.

By contrast with the single error that typically constitutes a bias, the random errors that occur during a measurement procedure are many. The collective effect of random errors is determined by replicate analysis of a *test specimen* (or internal quality control, IQC) and is characterised by a standard uncertainty. When the value obtained for the measurand is biased, there are two ways of proceeding:

1. To correct for the bias, effectively eliminating the bias from the estimate of the expanded uncertainty (but retaining the option to report the value of bias explicitly). This procedure entails the established and uncontentious statistical analysis that yields the value (s) of the measurand(s) and their expanded uncertainties.
2. To incorporate the bias itself within a now suitably enlarged value of expanded uncertainty.

The 'total error' (TE) or 'total analytical error' (TAE) model introduced by Westgard et al [3–5], and extensively discussed over a number of years [4–10], basically adds an expanded measurement uncertainty to an absolute bias value to establish an enlarged range of uncertainty:

$$\mathrm{TE} = |b| + zu,$$

where the factor z (or coverage factor k) has been defined for various analytical situations by Stöckl and Thienpont [18]: $z = 1.96$ for a two-sided 95% limit in a Gaussian PDF. Oosterhuis et al [7] have commented further on TE and related methods. Five more methods are summarised by Magnusson and Ellison [19]. A further method is described by Synek [22], as an elaboration of previous work by the same author [14]. In [22], a Gaussian PDF is calculated that peaks at the estimated value of bias. Using tables or an algorithm, two points on the horizontal axis are then calculated that are equidistant from the origin and between which 95% (or any other selected percentage) of the area of the Gaussian is enclosed. In addition, the *International Vocabulary of Metrology – basic and general concepts and associated terms* also indicates that 'sometimes estimated systematic effects are not corrected for but, instead, associated measurement uncertainty components are incorporated' [1, Note 1 to item 2.26]. In all these methods, the bias enters into the expression for U. In one case among the several examined in [11] (cited in [19]), the expanded range is asymmetric with respect to the estimated value of the measurand.

In Sections 2 and 3 we consider the general case where there may be more than one reference and where the test is compared with the reference using replicate test measurements. Even though bias, like MU, may vary with concentration (an example of 'variation in a predictable manner' [1]), we assume here that any bias is constant (or approximately constant) across the measurement range in order to simplify the statistics and associated discussion. If a bias that is proportional to concentration is observed, the measurement range can be divided into segments with the appropriate bias allocated to each segment (in a similar manner in which MU is required to be assessed at individual points within the measurement range).

Among the various procedures for incorporating an uncorrected bias, we examine the widely quoted and often implemented Nordtest procedure [23]. We believe that this procedure could be improved, and in particular we propose an alternative method of calculating bias uncertainty. In Section 2 we establish basic relationships which will then be used to outline some differences in approach for including uncorrected bias into the overall uncertainty statement. In Section 2.1 bias is formally defined as a metrological condition independent of any experimental scatter in the test results. Section 2.2 summarises some common laboratory procedures. Section 2.3 discusses the uncertainty of a biased test result and the consequent PDF of the test. Section 2.4 discusses the estimation of the uncertainty of bias in the general case of several references with replicate measurements. Sections 2.5 and 2.6 discuss the uncertainties of the test and reference respectively. Section 2.7 shows the connection between the material in Sections 2.4 and 2.5 with the GUM treatment of uncorrected bias.

In Section 3, we describe a new statistical approach for the correction of bias, using the formulas developed in Section 2, and we propose a graphical method that provides a visual representation of reference and test data. Using Gaussian PDFs for the reference and test data, the common overlap area of these two curves provides a measure of the concordance of the two data sets. This is further discussed in Section 4.

2. Methods for treating uncorrected bias

2.1. Definition of bias and the relationship between uncertainties of bias and of reference

Bias is usually understood as a persistent difference between the value (or the mean value) of the test result and the reference quantity. In general, there exists a reference with a certified value x_{ref} with known standard uncertainty u_{ref}. This is an implicit or Type B uncertainty. There also exists a test whose result x_{test} is compared with x_{ref}. That is, the reference material is analyzed by the test method and the result x_{test} is compared with the given value x_{ref} of the reference. The measurand x_{test} has an estimated standard uncertainty u_x, determined explicitly, for example from the scatter of replicate measurements as a Type A uncertainty, or (perhaps less commonly) implicitly from documented past experience as a Type B uncertainty [GUM [2], Sections 2.3.2, 2.3.3 and 3.3.5].

We now define bias by making use of the following observation: a bias b_0 and its standard uncertainty u_b are metrological states independent of the scatter of the test results, whether this scatter is explicit (Type A) or implicit (Type B). In order to express this independence, we define b_0 as:

$$b_0 = \lim_{N_{\text{test}}\to\infty} \bar{x}_{\text{test}} - x_{\text{ref}}, \tag{1}$$

where $\bar{x}_{\text{test}} = (1/N_{\text{test}})\sum_{j=1}^{N_{\text{test}}} x_{\text{test}j}$ is the mean of a practically obtainable and representative sample of size N_{test} of test results and $x_{\text{test}j}$ is the jth test measurement. We denote the variance of the mean $\bar{x}_{test}$ of the sample as $u_{\bar{x}}^2$ and denoting the variance of the bias b_0 as $u_b{}^2$, and assuming negligible mutual correlation of $\bar{x}_{\text{test}}$ and x_{ref}, Eq. (1) now gives

$$u_b^2 = u_{\bar{x}}^2 + u_{\text{ref}}^2, \tag{2}$$

with

$$u_{\bar{x}}^2 = \lim_{N_{\text{test}}\to\infty} \frac{1}{N_{\text{test}}(N_{\text{test}}-1)} \sum_{j=1}^{N_{\text{test}}} (x_{\text{test}j} - \bar{x}_{\text{test}})^2. \tag{3}$$

In Eq. (3), the factor $1/(N_{\text{test}} - 1)$ is the standard factor in the expression for estimating the variance of a population from the variance of a representative sample. The factor $1/(N_{\text{test}})$ arises because if N_{test} measurements (assumed to be uncorrelated) have variance u^2, then the variance of their mean is u^2/N_{test}. As N_{test} increases to infinity, $u_{\bar{x}}^2$ in Eq. (3) approaches zero. Eq. (2) therefore becomes

$$u_b = u_{\text{ref}}. \tag{4}$$

In effect, the bias has been defined as the difference between the reference and the mean of the test results where this mean has been assigned a negligible uncertainty, and is therefore effectively a fixed point estimate. Any uncertainty in the reference is then directly reflected in an equal uncertainty in the bias. Thus, since uncertainty is the collective effect of errors, the random scatter of test results is separate from any given error in the reference, whether this error is positive or negative. All the test results will then be shifted by an equal amount in the positive or negative direction respectively.

Eq. (4) will be applied to several examples in this review. However, we note that there may be cases where u_b is known from separately obtained information, and Eq. (4) may no longer apply to such cases.

2.2. Common analytical procedures

In laboratory medicine, the 'trueness' of a measured quantity may be assessed by various procedures which include: concordance with a commutable CRM, concordance with a reference measurement procedure or concordance with values provided by an external quality assessment program (EQA, or an external proficiency testing scheme). Whichever procedure is defined (or assumed) as yielding the 'true' value of the measurand, it is described in this article as the reference procedure or reference. In particular, the statistics presented describe a procedure for comparing a test result or an analytical test method with the chosen reference, but the procedure itself is independent of the 'true' value of the reference data.

A particular selected reference procedure will generally fall into one of the following categories:

1. The use of single commutable CRM analyzed once or in replicate.
2. The use of two or more commutable CRMs with different measurand values, each analyzed once or in replicate.
3. Comparison of a test method with a reference method using patient (or other commutable) samples. In this situation, there are likely to be multiple sample pairs with measurement values covering the working analytical range (analyzed by the test method and by the reference method).
4. The use of EQA samples and the assumption that the stated values (the target values, method means or method medians) represent the 'true' assigned measurand value(s). In this situation, it is again likely that there will be several comparisons at different measurement values.

In the discussion that follows, any of the above scenarios can be implemented to ascertain the trueness of the test method and a numerical assessment of any bias.

2.3. The uncertainty of a biased test result

In order to provide an alternative to the procedures described by Nordtest [23] and Westgard [3–5] (for example), equations for incorporating uncorrected bias using conditional probabilities have been derived previously [25]. In the absence of bias, the test has the measured value $x_{\text{test}} = x_{\text{ref}}$ with standard uncertainty u_x. We assume for the moment that there is only one reference and no replication of measurements. In the presence of bias b_0 with standard uncertainty u_b, the test yields the result $x_{\text{test}} = x_{\text{ref}} + b_0$ with standard uncertainty $\sqrt{u_x^2 + u_b^2}$. This root-sum-square can be called the 'combined' standard uncertainty, u_{comb}, so that

$$u_{\text{comb}} = \sqrt{u_x^2 + u_b^2}. \tag{5}$$

The PDF of the test will therefore have a mean $x_{\text{ref}} + b_0$ and standard uncertainty u_{comb} given by Eq. (5). Eq. (5) holds whether or not the bias has been corrected. As can be shown, Eq. (5) remains valid regardless of the actual PDFs of test or bias. These PDFs may be, for example, Gaussian, lognormal (logGaussian) or uniform, which may be pertinent to many areas of clinical research [7,23,27–30].

Eq. (5) will be used to determine the PDF of the test, in connection with a proposed *figure-of-merit* for test-reference comparisons to be introduced and discussed in Section 3.

2.4. Estimation of uncertainty of bias

The test is usually compared with the reference using replicate measurements. The scatter of the test results is then described by the standard uncertainty u_x. For a given reference, in fixed conditions (same concentration, temperature, equipment, environment and operator), any accompanying bias is independent of the number of replicate measurements. This is a general characteristic of systematic error or bias, which tends to 'resist statistical attack', in the sense that mere replication of measurements, under fixed conditions, reduces neither the observed bias nor its uncertainty.

Table 1
Meaning of symbols.

Symbol	Meaning
x_{ref}	Reported value of the reference
x_{test}	Value obtained for the reference as determined by the test method
u_x	Scatter, measured as a standard uncertainty, associated with x_{test}
u_{ref}	Known standard uncertainty of the value of the reference
b_0	Single value of bias, for only one reference and no replication of measurements
u_b	Standard uncertainty of the bias for a given reference; equated to u_{ref}
N	Number of references
n_1	Number of replicate measurements using Reference 1
n_2	Number of replicate measurements using Reference 2
n_i	Number of replicate measurements using Reference i
n_N	Number of replicate measurements using Reference N
M	Total number of replicate measurements: $M = n_1 + n_2 + \ldots + n_N$
$\bar{b}_1$	Mean bias over n_1 replicates, using Reference 1
$\bar{b}_2$	Mean bias over n_2 replicates, using Reference 2
$\bar{b}_i$	Mean bias over n_i replicates, using Reference i
$\bar{b}_N$	Mean bias over n_N replicates, using Reference N
$\bar{b}_w$	Weighted mean value of bias, to be used when n_i varies with i

We now assume N references associated with respective mean biases $\bar{b}_1$, $\bar{b}_2$, …,$\bar{b}_N$ and respective standard uncertainties of bias u_{b1}, u_{b2}, …,u_{bN}. These standard uncertainties, following Eq. (4) are implicit standard uncertainties equal to the respective standard uncertainties $u_{\text{ref }1}$, $u_{ref\ 2}$, …,$u_{ref\ N}$ of the references. For the ith reference, the number of replicate measurements using that reference is denoted n_i. Explicit variations over the n_i replicates, for the ith reference, are estimated as standard uncertainties u_{xi}, ($i = 1,2, \ldots ,N$).

Table 1 summarises the symbols and their meanings.

When there are $N > 1$ references, it is desirable to include the variation in bias over all N references as *also* contributing to the *total* uncertainty in the bias. Accordingly, we make use of the following statistical relationship that is of general scope but is here formulated specifically for the purpose of determining the uncertainty in the bias. As discussed further below, a distinction is made between the prefixes 'intra' and 'inter'.

$$\text{total variance of bias} = \text{weighted mean of } N \text{ intra-measurement variances of mean biases} \\ + \text{ weighted inter-measurement variance of } N \text{ mean biases.} \tag{6}$$

The weighting is necessary if the n_i vary with i ($i = 1, 2, \ldots, N$). This would imply that the measurements on the references are not all equal in number. An example occurs in Nordtest [23, page 20] where there are three CRMs ($N = 3$), but the first is used 12 times, the second 7 times and the third 10 times; thus $N = 3$ and $n_1 = 12$, $n_2 = 7$ and $n_3 = 10$. We denote by M the sum of all the n_i; thus $M = n_1 + n_2 + \ldots + n_N$ and in this Nordtest example $M = n_1 + n_2 + n_3 = 12 + 7 + 10 = 29$.

The weighted mean $\bar{b}_w$ of all the mean biases, over all N references, is given by

$$\bar{b}_w = \frac{1}{M}\sum_{i=1}^{N} n_i \bar{b}_i, \tag{7}$$

It may be checked that, if $n_1 = n_2 = n_3 = \ldots = n_N = n$, then $M = Nn$ and Eq. (7) reduces to

$$\bar{b} = \frac{1}{Nn} n \times \sum_{i=1}^{N} \bar{b}_i = \frac{1}{N}\sum_{i=1}^{N} \bar{b}_i, \tag{8}$$

namely the ordinary unweighted mean $\bar{b}$ of the N biases.

We now consider the two terms on the right-hand side of Eq. (6). The total variance of bias is split up in this way for the following reasons.

1. In the first term, the N intra-measurement variances of biases are the N values of u_{bi}^2 ($i = 1, 2, \ldots, N$). For each reference, say the ith, u_{bi}^2 is the square of the standard uncertainty (perhaps expressed as a multiple of this, namely as an expanded uncertainty) stated in (or inferred from) the report or certificate on the ith reference. Over all N references, a weighted mean of all such N variances is calculated.
2. The second term refers to the variance of the biases themselves. In the unweighted case, this variance is the general formula that calculates the variance of a sample of size m, consisting of readings y_1, y_2, …,y_m as:

$$\text{variance} = \frac{1}{m}\sum_{i=1}^{m} (y_i - \bar{y})^2, \tag{9}$$

$\bar{y}$ being the mean $\bar{y} = (1/m)\sum_{i=1}^{m} y_i$ of the sample. We note that the divisor in Eq. (9) is m, not $m - 1$; this latter divisor would be appropriate if the variance of the sample were to be used to estimate the variance of a much larger population. However, such is not the purpose here, since all the variances in Eq. (6) are variances of samples.

Eq. (9) is equivalent to: variance = mean square minus squared mean:

$$\text{variance} = \left(\frac{1}{m}\sum_{i=1}^{m} y_i^2\right) - \bar{y}^2. \tag{10}$$

This convenient formulation in Eq. (10) can be prone to round-off errors, so that extra decimal places are advisable when using Eq. (10), and appropriate rounding should be applied only to the final result.

The second term in Eq. (6) therefore refers to separate measurement sets that may involve different references or the same reference but used in varying circumstances, for example with different reagent lots, different operators or different calibrators. Thus the two terms on the right-hand side of Eq. (6) will tend to occur respectively over a short and a longer time-scale. We refer to the first term on the right-hand side of Eq. (6) as the weighted mean wMoV(bias) of the N 'intra-measurement' variances of bias. We refer to the (single) variance in the second term on the right-hand side of Eq. (6) as the weighted 'inter-measurement' variance wVoM(bias) of biases.

We have, therefore, for wMoV(bias) the expression

$$\text{wMoV(bias)} = \frac{1}{M}\sum_{i=1}^{N} n_i u_{bi}^2 \tag{11}$$

Again, it may be checked that if $n_1 = n_2 = n_3 = \ldots = n_N = n$, since we now have $M = Nn$ Eq. (11) gives $\text{wMoV(bias)} = \frac{1}{N}\sum_{i=1}^{N} u_{bi}^2$, which is simply the ordinary mean of the u_{bi}^2.

The term wVoM(bias) is given by the formula that is a generalisation of Eq. (10):

$$\text{wVoM(bias)} = \left(\frac{1}{M}\sum_{i=1}^{N} n_i (\bar{b}_i)^2\right) - (\bar{b}_w)^2. \tag{12}$$

If $n_1 = n_2 = n_3 = \ldots = n_N = n$, Eq. (12) reduces to $\text{wVoM(bias)} = \left(\frac{1}{N}\sum_{i=1}^{N} (\bar{b}_i)^2\right) - \bar{b}_w^2$, which is the ordinary variance of the $\bar{b}_i$, and $\bar{b}_w$ is now the ordinary mean $\bar{b}$ of the N values of b_i, $i = 1, 2, \ldots, N$: that is, the variance is now simply 'mean square minus squared mean'.

The total variance of bias, V_b(tot), following Eq. (6), is given by the sum

$$V_b(\text{tot}) = \text{wMoV(bias)} + \text{wVoM(bias)} \\ = \left(\frac{1}{M}\sum_{i=1}^{N} n_i u_{bi}^2\right) + \left(\frac{1}{M}\sum_{i=1}^{N} n_i (\bar{b}_i)^2\right) - (\bar{b}_w)^2. \tag{13}$$

We now set $u_b = u_{\text{ref}}$ as stated in Eq. (4), for each of the N references $i = 1, 2, \ldots, N$. Thus we have $u_{bi} = u_{\text{ref }i}$, for $i = 1, 2, \ldots, N$. The total variance is therefore

$$V_b(\text{tot}) = \left(\frac{1}{M}\sum_{i=1}^{N} n_i u^2_{\text{ref}\,i}\right) + \left(\frac{1}{M}\sum_{i=1}^{N} n_i(\bar{b}_i)^2\right) - (\bar{b}_w)^2, \tag{14}$$

and so the total standard uncertainty u_b(tot) of the bias is

$$u_b(\text{tot}) = \sqrt{V_b(\text{tot})}. \tag{15}$$

We note the different approach used for the three-CRM case in Nordtest [23, page 20], and also in [19] (Eqs. (3) and (4) in [19]). In both cases the standard uncertainty of the bias that results from its variability over the N CRM's (in our terminology $\sqrt{\text{wVoM(bias)}}$, a measure of the 'inter-measurement variability') is equated to the root-mean-square value of the actual biases. When $N = 1$, it is stated in [23] and in [19] that the standard uncertainty of the bias has a component equal to the actual bias. These approaches in [19,23] would be more statistically correct if the mean bias $\bar{b}$ was zero or insignificant as actually suggested by Magnusson and Ellison [19]. This standard uncertainty is more generally given by $\sqrt{\text{wVoM(bias)}}$, and this is zero if $N = 1$.

We note, further, that a bias which may appear to be insignificant, because of its own high uncertainty, will actually be significant if the same value of bias (or approximately the same value) is obtained after repeated measurements. Thus suppose that the bias is estimated at +1% with a standard uncertainty also 1%, and so the bias is considered to be insignificant. However, if the same value of bias, +1%, is obtained repeatedly, say after ten measurements, the mean bias will of course still be +1% but the standard uncertainty of this value will now be only about 0.32%, implying a significant bias.

2.5. Uncertainty of test results

The test results are the results obtained for the values of the references when determined using the test method. The test results are characterised by variances $u_{x1}{}^2$, $u_{x2}{}^2$, ...,$u_{xN}{}^2$, which are again assumed to be constant (or approximately so) across the particular relevant measurement range. Each corresponding standard uncertainty u_{xi} is assumed to be a measure of the scatter of the test results for the ith reference ($i = 1, 2, \ldots, N$). Thus the measure of the resulting scatter of the *mean* test result for the ith reference, based on n_i mutually uncorrelated measurements, is the standard uncertainty $u_{xi}/\sqrt{n_i}$, for $i = 1, 2, \ldots, N$. This accounts for the extra factor $1/n_i$ in the squared term $u_{xi}{}^2/n_i$ in Eq. (16) to follow. So the weighted mean wMoV(test) is given by the expression:

$$\text{wMoV(test)} = \frac{1}{M}\sum_{i=1}^{N} n_i(u^2_{xi}/n_i) = \frac{1}{M}\sum_{i=1}^{N} u^2_{xi}. \tag{16}$$

For the test results, there is no equivalent to an inter-measurement weighted variance of means, since the actual inter-measurement variability of the test results is not regarded as contributing to a total uncertainty in the way that inter-measurement variability in bias does.

The PDF of the overall test result has mean $x_{\text{ref}} + \bar{b}_w$ and standard uncertainty u_{comb} given by

$$u_{\text{comb}} = \sqrt{\text{wMoV(test)} + V_b(\text{tot})}, \tag{17}$$

a generalisation of Eq. (5).

2.6. Uncertainty of reference results

The weighted mean of the intra-measurement variances of the references wMoV(ref) is given by the formula identical to Eq. (11):

$$\text{wMoV(ref)} = \frac{1}{M}\sum_{i=1}^{N} n_i u^2_{\text{ref}\,i} \tag{18}$$

For a single reference, $N = 1$, Eq. (18) reduces to MoV(ref) = $u_{\text{ref}}{}^2$ and is the variance of the reference. For several references, $N > 1$, Eq. (18) determines the overall variance of the combination of N references. In either case the square root of wMoV(ref), a standard uncertainty, is needed to evaluate the figure-of-merit to be described in Section 3, namely the degree of overlap of reference and test PDFs.

The application of Eqs. (4)–(18) to particular cases will be illustrated in Section 3.

2.7. Consistency with recommendations in GUM F2.4.5

The treatment of bias in GUM [2, F2.4.5] is expressed in terms of integrals over a variable t, which could be (for example) a continuous variable representing the temperature for those cases where bias has a known dependence on temperature. The variable t could also represent, in a discrete manner, a succession of different CRMs, indexed in our terminology by the subscript i. In the GUM terminology, $y(t)$ may therefore be regarded as the mean of the test results for the CRM indexed by a particular value of t. So these integrals are essentially equivalent to (unweighted) summations as in Section 2.4 above. For example, Eq. F.7a in GUM is stated as follows, for calculating a mean bias $\bar{b}$:

$$\bar{b} = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} b(t)dt, \tag{19}$$

but this is essentially the same as

$$\bar{b} = \frac{1}{N}\sum_{i=1}^{N} \bar{b}_i, \tag{20}$$

as in Eq. (8).

Similarly, Eq. F.7b in GUM for $u^2(\bar{b})$ is an unweighted wVoM, or VoM, and Eq. F.7c in GUM for $\overline{u^2(b(t))}$ is the MoV. As in Eq. (13), these two are summed in Eq. F.7e in GUM.

For a test result $y(t)$ that is subject to an uncorrected bias $b(t)$, and also to a standard uncertainty $u(y(t))$ from all sources of uncertainty other than bias, equ. F.7e in GUM [2, F2.4.5] gives the variance of the test result $y(t)$ as $\overline{u^2(y(t))} + \overline{u^2(b(t))} + u^2(\bar{b})$. This sum, in our terminology and with the three summed terms in the same order, is wMoV (test) + wMoV(bias) + wVoM(bias). For a 95% confidence interval with a large number of degrees of freedom, the coverage factor k is $k = 2$ and the expanded uncertainty U attaching to $y(t)$ is then $U = k\sqrt{\overline{u^2(y(t))} + \overline{u^2(b(t))} + u^2(\bar{b})}$. This is equivalent to (in our terminology) Eq. (17) multiplied by k.

3. A proposed method for the treatment of bias

A general feature of all proposals for incorporating uncorrected bias is that a single bias, or its square, is summed with uncertainty components. We should note, however, a fundamental distinction between bias and uncertainty. An uncertainty arises as the collective effect of *unobservable* random errors. If these errors could be observed, and therefore individually taken into account, the uncertainty would no longer exist. In contrast, a bias is an *observable* error (even though it might not be immediately observed, confirmed, and quantified). As a consequence of bias being an observable error, an estimate can be obtained of its numerical value, its sign will be known, and the standard uncertainty of the bias, equal to the known standard uncertainty of the reference, is also known. We therefore propose that since the bias is an observable and quantified error with a known sign and standard uncertainty, a correction for the bias can usually be made. The standard uncertainty of the bias will enter into calculating the standard uncertainty of the test and hence its expanded uncertainty. In reporting the results of the measurement, the value of bias and its standard uncertainty can be stated separately, if so desired. Correcting for bias would avoid any asymmetric expanded uncertainty range.

The proposed method for the treatment of bias is based on Eq. (5). Eq. (5), multiplied by a coverage factor k so as to provide an expanded uncertainty U, appears in the 'correction applied' Section of Table 3 in

[19] and also appears in GUM [2, Eq. F7e], in [22] and in Eurachem [26].

As discussed in [25], a figure-of-merit may be devised to indicate the quality of the test, where 'quality' indicates the measured degree of closeness of the test to the reference. More specifically, a graph is drawn of PDF(ref) of the reference, and on the same axes the PDF(test) of the test. In Figs. 2 and 3, the horizontal axis, shared by the reference and test PDFs, will be labeled 'Measurement value', applying to either PDF. The quality, and therefore the figure-of-merit, is then indicated by the percentage overlap of PDF(ref) and PDF(test). A simple algorithm for calculating the overlap is described in [25] and an Excel (Microsoft Corporation, Redmond, WA, USA) procedure for this algorithm is provided in the Supplementary Appendix. It may be generally assumed that both PDF's are Gaussian density distributions. With very small sample sizes, say $n = 6$ or less, the t-distribution on $n - 1$ degrees of freedom may be preferable. (As observed below in connection with Example 2, with 12 observations the density distribution is virtually indistinguishable from a Gaussian). The graph PDF(ref) has mean x_{ref} and standard uncertainty u_{ref}. The graph PDF(test) has mean $x_{ref} + b_0$ and standard uncertainty $u_{comb} = \sqrt{(u_x^2/n) + u_b^2} = \sqrt{(u_x^2/n) + u_{\text{ref}}^2}$, where u_x is the standard uncertainty of n assumed uncorrelated test results and so $u_x/\sqrt{n}$ is the standard uncertainty of their mean.

Three examples, an artificial and two real-world examples, will now be analyzed in order to illustrate the proposed process of bias correction. The real-world examples are the same as those described in Nordtest [23, page 20] and any differences from the original treatment will be noted.

Example 1.

Fig. 1 shows the procedure by which the correction for bias may be carried out. Fig. 1A shows a situation without bias, so that the mean of PDF(test) coincides with the mean of PDF(ref). Both PDF's are assumed to be Gaussian. An arbitrary standard uncertainty $u_{\text{ref}} = 0.5$ is assigned to PDF(ref) and an arbitrary standard uncertainty $u_x = 1$ to PDF(test). Since the total area under a PDF must always be 1, the height of the test curve is one-half that of the reference curve. The overlap of reference and test, for zero bias, is 67.7%. (We state three decimal places for overlap percentages, in order that readers may, if they wish, check their overlap calculations against the algorithm provided in the Supplementary Appendix). Fig. 1B shows the situation where a bias b_0 exists and the arbitrarily assigned values for the bias and for its standard uncertainty are $b_0 = + 0.8$ and $u_b = 0.5$. The peaks of the two curves are now separated by $b_0 = + 0.8$ and the combined standard uncertainty of PDF(test) is $u_{\text{comb}} = \sqrt{u_{\text{ref}}^2 + u_x^2} = \sqrt{0.5^2 + 1.0^2} = 1.12$. The overlap of PDF(ref) and PDF(test) is 51.2%. Fig. 1C shows the correction for bias, which moves PDF(test) back to a mean value of 0, coinciding with the mean value of PDF(ref), and with 1.12 as the standard uncertainty of the bias-corrected test. (The 'post-correction' overlap with PDF(ref), which could also be a useful calculation, is in this particular case 62.9%). The expanded uncertainty, taking a coverage factor $k = 2$ for a 95% confidence range, is $U = \pm 2.2$. This range is indicated as the portion of the horizontal axis enclosed by the two tick-marks in Fig. 1C.

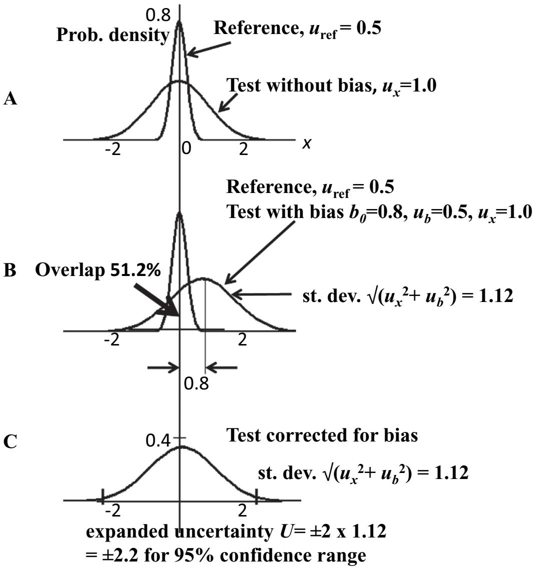

Fig. 1. Procedure for bias correction. See Example 1 in Section 3.

Example 2.

This and the next example are both based on data as described in Nordtest [23, page 20] and use proportional units expressed as percentages, that is, relative standard uncertainties.

In the first case, a single CRM with certified value $x_{\text{ref}} = 11.5 \pm 0.5$ for a 95% confidence interval (so $k = 2$) is considered. The standard uncertainty of the value 11.5 is therefore $u_{\text{ref}} = 0.25$, or u_{ref} (rel) = 0.25/11.5 = 2.16% (the percentage value stated in Nordtest [23]). In what follows it is understood that standard and expanded uncertainties and biases described as percentages are *relative* standard and expanded uncertainties and biases, respectively.

There are $n = 12$ test results with mean 11.9 and a standard uncertainty of 2.2%. The standard uncertainty of the *mean* of the 12 (assumed mutually uncorrelated) results is u_x (rel)/$\sqrt{12}$ = 2.2%/$\sqrt{12}$ = 0.635%. The bias is calculated as (11.9 − 11.5)/11.5 = + 3.48%. If $u_{\text{ref}} = 0.25$, then we assume $u_b = 0.25$ also, and so as a percentage, referring u_b to the actual test value 11.9, u_b(rel) = 0.25/11.9 = 2.10%. Then u_{comb} (rel) = $\sqrt{\left(\frac{2.2}{\sqrt{12}}\right)^2 + 2.10^2} = \sqrt{0.635^2 + 2.10^2}$ = 2.19%. It may be checked that these results agree with Eqs. (16) and (17) when we set $N = 1$, $n_1 = 12$ and $M = 12$.

Fig. 2A shows the situation before bias correction, relative to a Gaussian PDF(ref) with mean $x_{\text{ref}} = 0.0$ and standard uncertainty u_{ref}(rel) = 2.16%. The PDF(test) has a mean that is shifted right by the amount of bias +3.48% and a standard uncertainty 2.19%. In Fig. 2A, PDF(test) is drawn as a Gaussian PDF and the overlap is 42.4%. If PDF (test) was drawn more accurately as a scaled and shifted t-distribution on 12 − 1 = 11 degrees of freedom, PDF(test) would be virtually indistinguishable from a Gaussian and the overlap would be slightly less by an insignificant ≈1%. (A Gaussian PDF is essentially a t-distribution on a very large number of degrees of freedom). Because each PDF must have a total area of 1, and because PDF(test) has the slightly greater standard uncertainty, PDF(test) is slightly vertically shorter than PDF (ref). It is clear that if the bias were smaller, the overlap area would be greater.

For the test in this example we take a coverage factor $k = 2.2$, appropriate to 12 − 1 = 11 degrees of freedom for a 95% confidence range associated with this t-distribution. So the expanded uncertainty U (rel) is given by $U(\text{rel}) = \pm ku_{\text{comb}} = \pm 2.2u_{\text{comb}} = \pm 2.2 \times 2.19\,\% \approx \pm 4.82\%$. Since the observed mean of the test is 11.9, U in absolute units is (4.82/100) × 11.9 = 0.57. The bias-corrected report on the above example might run as follows, where the relevant units are in square brackets as [units].

The mean value of the test is estimated to be 11.5 [units] with an expanded uncertainty U given by $U = \pm 0.57$ [units] for a 95% confidence range. There is therefore 95% probability that the value of the analyte lies between 11.5 − 0.57 = 10.9 and 11.5 + 0.57 = 12.1 A bias of +0.4 [units] was observed, having a standard uncertainty of 0.25 [units]. The stated mean value of the test has been corrected for this bias.

We contrast the discussion above with the approach in Nordtest [23, page 20] for a single CRM:

1. The standard uncertainty of 2.2% of the 12 test results is there given the label s_{bias}, but this is the *scatter* of the test results expressed as a

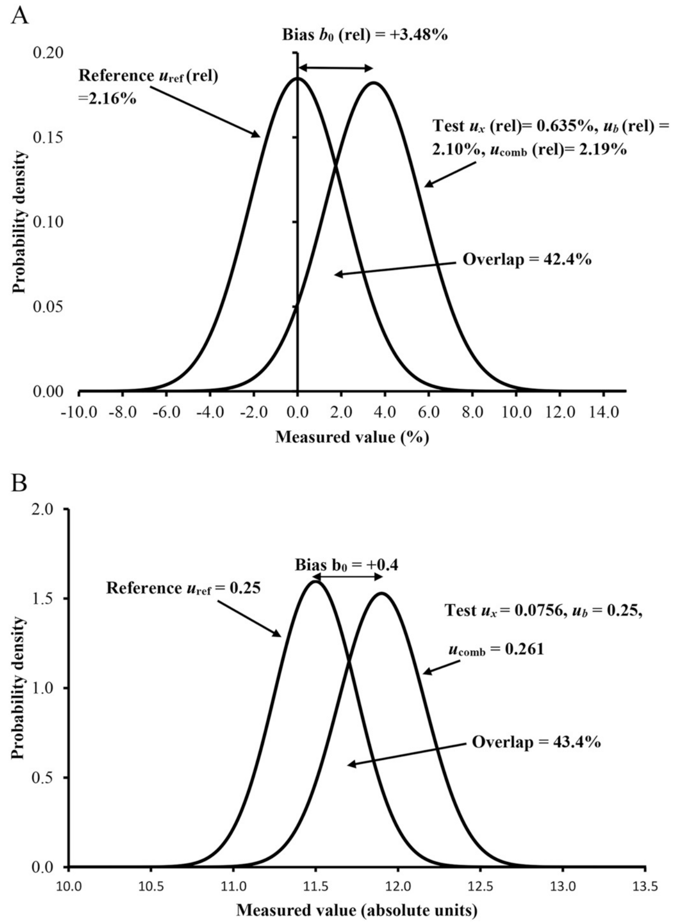

Fig. 2. A. Procedure applied to a real-world case of a single reference, before bias correction. Percentage values (see text). B. Procedure applied to a real-world case of a single reference, before bias correction. Absolute values. For both Figs. A and B, see Example 2 in Section 3.

standard uncertainty and in our terminology is u_x(rel). The quantity called u(*bias*), with the same meaning as u_{comb} in our terminology, is there equated to

$$u(bias) = \sqrt{3.48^2 + \left(\frac{2.2}{\sqrt{12}}\right)^2 + 2.16^2} = 4.1\%,$$

where 3.48% is the actual bias, 2.2% is the standard uncertainty of the test results, and 2.16% is the standard uncertainty of the reference. However, with a single CRM and a single estimate of bias, VoM(bias) = 0 as stated previously (after Eq. (15)), and the first term in u(*bias*), namely 3.48%, should be absent. This would give u (*bias*) = u_{comb}(rel) = 2.25% instead of 4.1%.

2. As a minor point of difference, our value for u_{comb}(rel) is 2.19% instead of the just-mentioned 2.25%, because we have referred u_b = 0.25 to the mean 11.9 of the test results, giving u_b(rel) = 0.25/ 11.9 = 2.10%. The reason is that the mean test result of 11.9 reveals the bias. In Nordtest, the value u_b = 0.25 is referred to the reference value 11.5, giving the ratio 2.16%.

Fig. 2B shows the same data as Fig. 2A but now in absolute, instead of proportional (percentage) units. The value u_x = 0.0756 is obtained as 0.635 % × 11.9. The overlap at 43.4% in Fig. 2B is not significantly different from the 42.4% of Fig. 2A. Percentage and absolute units give very closely similar overlaps (any small discrepancies arise from approximations involved in converting absolute values to percentages and from round-off errors). The reason is that overlap has an invariance property to be briefly discussed in Section 4.

Table 2
Data as described by Nordtest [23, page 20, three-CRM case], used for calculation of bias. The second-last column, headed 'u_x(rel)' gives the relative standard uncertainties that describe the amount of scatter of results in each of the three cases. The last column gives the resulting standard uncertainties of the means of the three measurements of scatter. These are denoted $u_x(\mathrm{rel})/\sqrt{n}$. The division by the square root of the number of measurements implies that the individual measurements have been assumed to be mutually uncorrelated.

CRM no.	n	Bias %	u_{ref}(rel)%	u_b(rel)%	u_x(rel) %	$u_x(\mathrm{rel})/\sqrt{n}$ %
1	12	+3.48	2.16	2.16	2.2	$2.2/\sqrt{12}$
2	7	−0.9	1.8	1.8	2.0	$2.0/\sqrt{7}$
3	10	+2.5	1.8	1.8	2.8	$2.8/\sqrt{10}$

Example 3.

This second example as discussed in Nordtest [23, page 20] involves three references. The results are summarised in Table 2. Again, biases and uncertainties are stated in relative terms as percentages, in conformance with the practice in [23].

We obtain from this table, and attaching 'rel' to denote relative values:

N = 3 (the number of CRM's).

M = 12 + 7 + 10 = 29 (the total number of replicate measurements, over all CRM's).

wMoV(ref)(rel) = 3.830 (from Eq. (18)), so $u_{ref}\,(\mathrm{rel}) = \sqrt{3.830} = 1.96\%$.

$\bar{b}_w\,(\mathrm{rel})$ = 2.085% (from Eq. (7)).

wMoV(bias)(rel) = 3.830 (from Eq. (11)).

wVoM(bias)(rel) = 3.015 (from Eq. (12)).

V_b(tot)(rel) = 6.845 (from Eqs. (13) or (14)).

u_b(tot)(rel) = 2.616% (from Eq. (15)).

wMoV(test)(rel) = 0.575 (from Eq. (16)).

$u_{comb}\,(\mathrm{rel}) = \sqrt{\mathrm{wMoV(test)(rel)} + V_b(\mathrm{tot})(\mathrm{rel})} = 2.724\%$ (from Eq. (17)).

Fig. 3 illustrates the reference and test PDFs before bias correction. In Fig. 3, the effective relative standard uncertainty of the combination of three references is indicated as u_{ref}(rel) = 1.96%. The value of u_{comb}(rel) for the combined three-CRM test is 2.72%, also to three decimal places. The reference-test overlap is 63.6%. This higher overlap and therefore better result than in Example 2 is largely the effect of a smaller overall bias.

We again contrast these results with the approach in Nordtest [23]:

1. The information about intra-measurement scatter (summarised in the second-last and last columns in Table 2 in this paper) does not

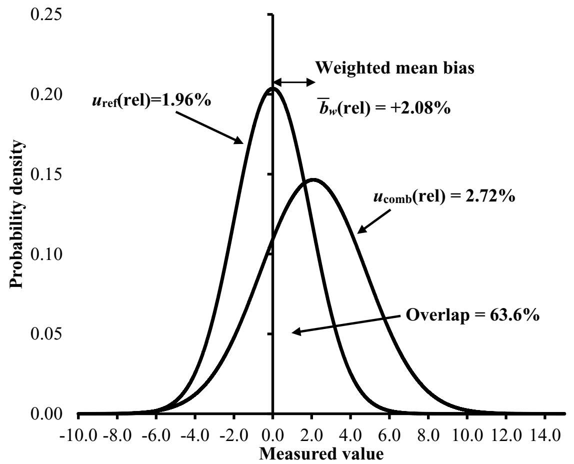

Fig. 3. Procedure applied to a real-world case of three references, before bias correction. See Example 3 in Section 3.

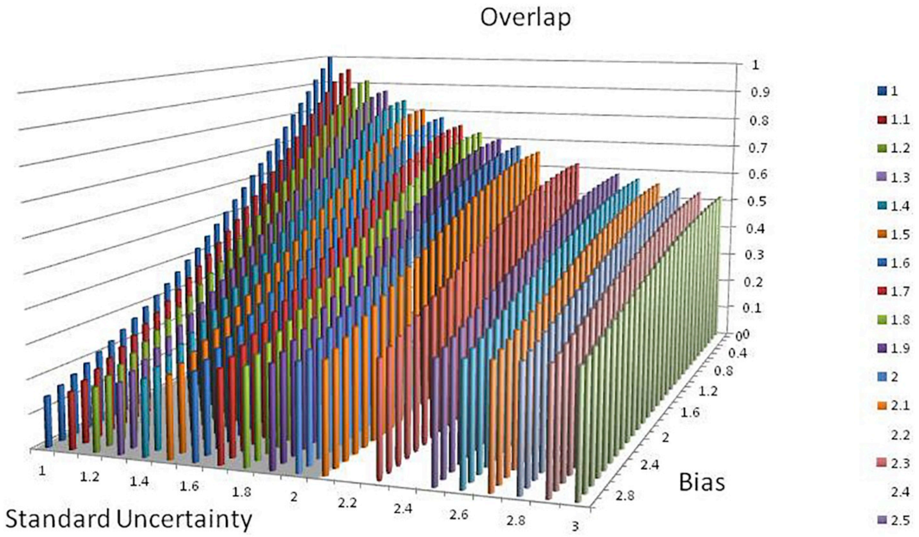

Fig. 4. Three-dimensional graph of overlap along vertical axis, and bias b_0 and combined standard uncertainty u_{comb} along horizontal axes.

appear to be used. In our approach, the intra-measurement scatter is $\sqrt{\text{wMoV(test)(rel)}} = \sqrt{0.575} = 0.76\%$.

2. There is no weighting of the results in Nordtest [23] to take into account the unequal number of measurements for the three CRMs. In [23, page 19, footnote 5] it is indeed suggested that weighting would be preferable. If the results are correctly weighted, the required total variance of the bias is the sum of the intra-measurement variances of bias and the inter-measurement variance of bias, as described in Section 2.4, but if there is no weighting for an unequal number of measurements this formula cannot be applied. As a consequence, the total variance of the bias cannot be correctly calculated.
3. The value called RMS_{bias} in Nordtest [23] is equated (without weighting) to $\sqrt{\frac{1}{3}(3.48^2 + 0.9^2 + 2.5^2)} = 2.53\%$, but this amounts to setting the mean bias as zero in Eqs. (7) or (8). However, in fact $\bar{b}_w(\text{rel}) = 2.085\%$ as stated above. Thus $RMS_{bias}{}^2 = 2.53^2 = 6.40$ using the approach in [23], but the correct value is wVoM(bias)(rel) in our notation, and this is 3.02 (to three decimal places).
4. In [23], a final value called $u(bias)$ is obtained as $u(bias) = 3.2\%$. This is the root-sum-square of 2.53% and 1.92%, where 1.92% is the unweighted mean of the three values of reference standard uncertainties, namely 2.16%, 1.8% and 1.8%. (In general, uncertainties should be combined by root-sum-squares rather than by simple linear addition, although admittedly in this particular case it makes little difference which procedure is used). In our approach, the final value $u_b(\text{tot})(\text{rel}) = 2.62\%$, and the resulting value of $u_{\text{comb}}(\text{rel})$ is 2.72%. This value $u_{\text{comb}}(\text{rel}) = 2.72\%$ is the standard uncertainty of the Gaussian representing the net measurement result over the three CRMs. See Fig. 3.

4. Some further comments on overlap

An interpretation of overlap is that it measures the probability that is common to both reference and test. Ideally, with zero scatter in the test result and zero bias, the common probability would be 1 (100%), implying identical probabilities for both reference and test. The actual overlap is a measure of their mutual disparity. In [25] we compared overlap with other figures-of-merit such as the E_n number, the Sigma-Metric, TAE and similar formulations.

Overlap has the useful property that it is invariant with respect to the same change of scale for u_{ref}, u_{comb} and bias b_0. Thus the same overlap applies to the case characterised by u_{ref}, u_{comb} and b_0 as to the case cu_{ref}, cu_{comb} and cb_0, where c is any positive number. Clearly the same overlap is obtained with negative bias as with the same value of positive bias. Overlap is also independent of the origin of measurements. In consequence, as Example 2 in Section 3 indicates, a closely similar overlap is obtained with proportional units, for example percentages, as with absolute units. More generally, the overlap is independent of the units of measurement provided that all terms are measured in the same units.

Fig. 4 is a three-dimensional graph with overlap as the vertical axis, and b_0 and u_{comb} as the horizontal axes. Gaussian PDFs are assumed for both PDF(ref) and PDF(test). The overlap is calculated with PDF(ref) having mean zero and standard uncertainty 1. In view of the just-stated invariance, Fig. 4 is generally applicable to any particular case provided that the two horizontal axes are appropriately scaled. Thus if a particular case has PDF(ref) having standard uncertainty u_{ref}, Fig. 4 describes the case with $c = 1/u_{\text{ref}}$. Then PDF(ref) has standard uncertainty $u_{\text{ref}}/u_{\text{ref}} = 1$, as appropriate to Fig. 4, and b_0 is scaled to b_0/u_{ref} and u_{comb} is scaled to $u_{\text{comb}}/u_{\text{ref}}$.

Fig. 5 is a contour plot corresponding to Fig. 4. The bands are separated by 0.05 (5%) overlap. Increasing bias always decreases the overlap. This can be seen from Fig. 5, where the horizontal dotted line is an arbitrary constant value of standard uncertainty (about 2.2 in Fig. 5). This horizontal line traverses the bands from left to right, and therefore in the direction of increasing bias, showing decreasing overlap as successive bands are traversed. This is so for any vertical position of the line and therefore for any value of standard uncertainty.

For the complementary case of constant bias and increasing standard uncertainty, the vertical dotted line nearest the left edge of Fig. 5 is situated at the constant low value of bias of 0.5. This vertical line traverses the bands upwards in the direction of increasing standard uncertainty, and it can be seen that the overlap decreases. This again is a desirable feature of overlap as a figure-of-merit. The central vertical dotted line in Fig. 5 is situated at the higher value of bias of 1.6. Because of the convex curvature of the bands in this region, it is seen that the overlap increases, to a value of about 52%, and then decreases again. The vertical dotted line nearest the right edge of Fig. 6 is situated at the high value of bias of 2.7 and here the overlap increases as the standard uncertainty increases. However, the feature of increasing overlap with increasing standard uncertainty occurs in any case at low values of overlap, between about 17% (at standard uncertainty 1.0) and 37% (at standard uncertainty 3.0) for a bias of 2.7.

Ideally, of course, both bias and standard uncertainty will be low. In practice, one or the other, or both, may well be high. As observed by Oosterhuis etal [7] in connection with TAE, the same TAE value may result from high bias and low standard uncertainty, as from low bias

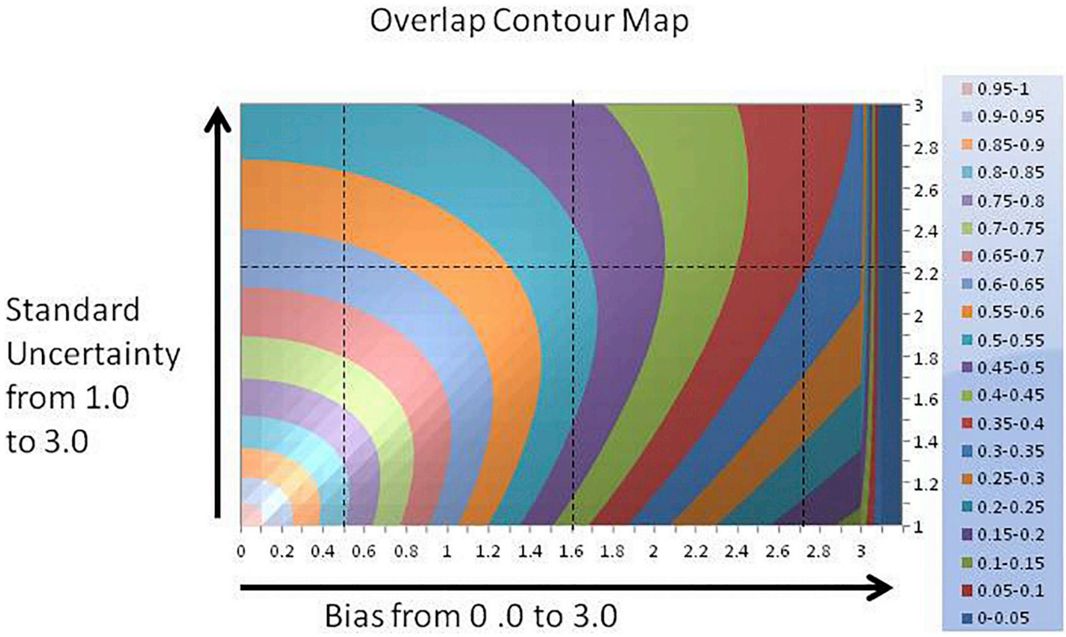

Fig. 5. Contour plot corresponding to Fig. 4.

and high standard uncertainty. This cannot occur when overlap is used as a figure of merit. As can be seen from Figs. 4 and 5, the combination of low bias and high standard uncertainty always results in a much greater overlap, and therefore indicates a better result, than high bias and low standard uncertainty. This is in accord with the general view that the latter case is the more serious problem.

5. Conclusion

Procedures for including uncorrected bias into the overall uncertainty statement for medical laboratory analyses have been reviewed. While GUM prefers all bias to be eliminated or corrected, the practical aspects of applied analytical chemistry allow an expanded uncertainty range which incorporates both systematic and random uncertainty components. The contentious nature of establishing such an expanded uncertainty range is demonstrated by the numerous articles describing bias and procedures for its inclusion into an expanded uncertainty. When bias and its uncertainty are recognised as metrological states independent of scatter in the test results, the uncertainty of the reference and uncertainty of the bias can be equated. The net standard uncertainty of the test results, which we call the 'combined' standard uncertainty, is the root-sum-square of the standard uncertainty of the bias and the standard uncertainty that is a measure of the scatter in the test results. Since an incomplete and therefore potentially erroneous formula is often used for estimating bias standard uncertainty, we propose an alternative calculation which is equivalent to the GUM F2.4.5 procedure. In addition, a graphical method is proposed which quantifies the discrepancy between the results of a test measurement, before bias correction, and the corresponding reference value in terms of the percentage overlap of two PDFs. A simple algorithm for calculating overlap using Excel is presented in the Supplementary Appendix. To demonstrate the proposed procedures and graphical method, we describe three typical examples and show how the calculation of overlap and a graphical approach may be of benefit, whether or not a correction for bias is made. However, we recommend that a correction for bias should be made wherever possible.

The recommendations in GUM F2.4.5 concern uncorrected bias. However, the GUM also recommends that bias should be corrected wherever possible (GUM 3.2.4), and although there are several GUM-related documents currently in preparation [31], it is likely that this recommendation will remain and apply to all areas of metrology. This is so after full acknowledgment of the particular metrological circumstances in analytical and clinical chemistry, with their much more complex and often less stable materials and greater environmental sensitivities than, for example, in physical metrology. It is helpful to think of metrology in general as the 'infra-structure' of the sciences, since all the sciences ultimately rely on accurate measurement with all systematic errors not only identified but also corrected for. We contend that such an approach should be attempted in analytical and clinical chemistry more frequently than now appears to be the case. As the importance of MU becomes more broadly accepted, the need to standardise its evaluation using the GUM framework is critical for ensuring ongoing consistency.

Competing interests

None.

Funding

None.

Ethical approval

Not applicable.

Guarantor

RF

Contributorship

All authors reviewed and edited the manuscript and approved the final version for publication.

Acknowledgements

We thank the reviewers of this article for their most useful and helpful comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cca.2019.03.1633.

References

[1] Joint Committee for Guides in Metrology, International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM), JCGM 200:2012. https://www.bipm.org/en/publications/guides/vim.html (Accessed December 2018).

[2] Bureau International Des Poids et Mesures, Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement (GUM), JCGM 100:2008, https://www.bipm.org/en/publications/guides/gum.html (Accessed December 2018).

[3] J.O. Westgard, R.N. Carey, S. Wold, Criteria for judging precision and accuracy in method development and evaluation, Clin. Chem. 20 (1974) 825–833.

[4] J.O. Westgard, S. Westgard, Quality control review: implementing a scientifically based quality control system, Ann. Clin. Biochem. 53 (2016) 32–50.

[5] J.O. Westgard, Useful measures and models for analytical quality management in medical laboratories, Clin. Chem. Lab. Med. 54 (2016) 223–233.

[6] E. Theodorsson, Uncertainty in measurement and total error: tools for coping with diagnostic uncertainty, Clin. Lab. Med. 37 (2017) 15–34.

[7] W.P. Oosterhuis, H. Bayat, D. Armbruster, A. Coskun, K.P. Freeman, A. Kallner, et al., The use of error and uncertainty methods in the medical laboratory, Clin. Chem. Lab. Med. 56 (2018) 209–219.

[8] J.O. Westgard, Error methods are more practical, but uncertainty methods may still be preferred, Clin. Chem. 64 (2018) 636–638.

[9] I. Farrance, T. Badrick, K.A. Sikaris, Uncertainty in measurement and total error – are they so incompatible? Clin. Chem. Lab. Med. 54 (2016) 1309–1311.

[10] I. Farrance, T. Badrick, R. Frenkel, Uncertainty in measurement and total error: different roads to the same quality destination? Clin. Chem. Lab. Med. 56 (2018) 2010–2014.

[11] S.J. Phillips, K.R. Eberhardt, B. Parry, Guidelines for expressing the uncertainty of measurement results containing uncorrected bias, J. Res. Natl. Inst. Stand. Technol. 102 (1997) 577–585.

[12] W. Hasselbarth, Measurement uncertainty procedures revisited: direct determination of uncertainty and bias handling, Accred. Qual. Assur. 3 (1998) 418–422.

[13] A. Maroto, R. Boqué, J. Riu, F.X. Rius, Should non-significant bias be included in the uncertainty budget? Accred. Qual. Assur. 7 (2002) 90–94.

[14] V. Synek, Attempts to include uncorrected bias in the measurement uncertainty, Talanta 65 (2005) 829–837.

[15] G.E. Donnell, D.B. Hibbert, Treatment of bias in estimating measurement uncertainty, Analyst 130 (2005) 721–729.

[16] D.B. Hibbert, Systematic errors in analytical measurement results, J. Chromat. A 1158 (2007) 25–32.

[17] L. Thienpont, Calculation of measurement uncertainty – why bias should be treated separately, Clin. Chem. 54 (2008) 1587–1588.

[18] D. Stöckl, L. Thienpont, About the z-multiplier in total error calculations, Clin. Chem. Lab. Med. 46 (2008) 1648–1649.

[19] B. Magnusson, S.L. Ellison, Treatment of uncorrected measurement bias in uncertainty estimation for chemical measurements, Anal. Bioanal. Chem. 390 (2008) 201–213.

[20] E. Theodorsson, B. Magnusson, I. Leito, Bias in clinical chemistry, Bioanalysis 6 (2014) 2855–2875.

[21] A. Kallner, 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. 54 (2016) 1291–1297.

[22] V. Synek, How to encompass an uncorrected bias into the expanded uncertainty with a fixed coverage probability: calculation procedures, Accred. Qual. Assur. 22 (2017) 179–186.

[23] B. Magnusson, T. Näykki, H. Hovind, M. Krysell, E. Sahlin, Handbook for Calculation of Measurement Uncertainty in Environmental Laboratories, Nordtest report TR 537 (edition 4) (2017).

[24] Eurachem/CITAC Measurement Uncertainty Traceability Working Group, Treatment of an Observed Bias, https://www.eurachem.org/index.php/publications/leaflets/bias-trt-01, May 2017 (Accessed December 2018).

[25] R.B. Frenkel, I. Farrance, A statistical procedure for the assessment of bias in analytical methods using conditional probabilities, Accred. Qual. Assur. 22 (2017) 265–273.

[26] Eurachem/CITAC and Guide CG4, Quantifying Uncertainty in Analytical Measurement, third edition, 2012, https://www.eurachem.org/index.php/publications/guides/quam#translations (Accessed December 2018).

[27] H.J. Møller, P.H. Petersen, L. Rejnmark, S.K. Moestrup, Biological variation of soluble CD163, Scand. J. Clin. Lab. Invest. 63 (2003) 15–21.

[28] M.R. Fokkema, Z. Herrman, F. Muskiet, J. Moecks, Reference change values for natriuretic peptides revisited, Clin. Chem. 52 (2006) 1602–1603.

[29] C. Klersy, G.V.M. D'Eril, A. Barassi, G. Palladini, M. Cornelli, R. Moratti, et al., Advantages of the lognormal approach to determining reference change values for N-terminal propeptide B-type natriuretic peptide, Clin. Chim. Acta 413 (2012) 544–547.

[30] P.H. Petersen, F. Lund, C.G. Fraser, G. Sölétormos, Analytical performance specifications for changes in assay bias (Δ Bias) for data with logarithmic distributions as assessed by effects on reference change values, Ann. Clin. Biochem. 53 (2016) 686–691.

[31] News from JCGM, June 2017, https://www.bipm.org/utils/en/pdf/news_jcgm-wg1.pdf (Accessed March 2019).