Counterpoint The *Guide to Expression of Uncertainty in Measurement* Approach for Estimating Uncertainty: An Appraisal

Jesper Kristiansen

Background: **The aim of the *Guide to Expression of Uncertainty in Measurement* (GUM) is to harmonize the different practices for estimating and reporting uncertainty of measurement. Although there are clear advantages in having a common approach for evaluating uncertainty, application of the GUM approach to chemistry measurements is not straightforward. In the above commentary, Krouwer suggests that the GUM approach should not be applied to diagnostic assays, because (*a*) the quality of diagnostic assays is to low, and (*b*) the GUM uncertainty intervals are too narrow to predict the outliers that occasionally trouble these methods.**

Methods: **Some of the examples presented by Krouwer are reviewed. Sodium measurements are modeled mathematically to illustrate the GUM approach to uncertainty. A standardized uncertainty evaluation process is presented.**

Results: **Modeling of sodium measurements demonstrates how the GUM uncertainty interval reflects the treatment of a bias: The width of the uncertainty interval varied depending on whether a correction for a calibrator lot bias was applied, but in both cases it was consistent with the distribution of measurement results. Expanding the uncertainty interval to include outliers runs counter to the definition of uncertainty. Used appropriately, the GUM uncertainty can be helpful in detecting outliers. In standardizing the uncertainty evaluation, the importance of the analytical imprecision and traceability was emphasized. It is problematic that manufacturers of commercial assays rarely inform about the uncertainty of the values assigned to the calibrators. As demonstrated by an example, external quality-assurance data may be used to estimate this uncertainty.**

Conclusions: **The GUM uncertainty should be applied to measurements in laboratory medicine because it may actually support the forces that drive the work on improving the quality of measurement procedures. However, it is important that the GUM approach is made more manageable by standardizing the uncertainty evaluation procedure as much as possible. It is essential to focus on the traceability and uncertainty of calibrators and reagents supplied by manufacturers of assays. Information about uncertainty is necessary in the evaluation of the uncertainty associated with manufacturers' measurement procedures, and in general it may force manufacturers to increase their efforts in improving the metrologic and analytical quality of their products.**

In the above commentary Krouwer *(1)* criticizes the so-called *Guide to Expression of Uncertainty in Measurement* (GUM)[1] uncertainty and concludes that although GUM uncertainty may be suitable for values assigned to reference materials, the application of the GUM approach to commercial diagnostic assays is not warranted. By dismissing GUM uncertainty, however, one misses the opportunity to use it as a tool to solve some of the measurement problems that riddle the field of laboratory medicine. In this commentary I will certainly not ignore the practical problems in applying the GUM approach, but in contrast to Krouwer, I see them as challenges that need to be solved.

Because much of the discussion presented here will concern properties of measurements (and results of measurements), I find it appropriate to start by considering what a measurement actually is. A measurement is defined in the metrologic vocabulary as a set of operations having the object of determining the value of a quantity *(2)*. The value of the quantity of interest (i.e., the measurand; for example, the concentration of glucose in blood) is estimated using a measurement procedure. When further pursuing the concept of measurement, one discovers that an overwhelming part of measurement procedures in analytical chemistry, and indeed in laboratory medicine as well, involves calibration. Measurements based on calibration are, in essence, comparisons. In other words, the measurement procedure is used to compare the pa-

The National Institute of Occupational Health, Lersø Parkallé 105, DK-2100 Copenhagen, Denmark. Fax 45-39-165201; e-mail jkr@ami.dk.

Received June 6, 2003; accepted August 11, 2003.
Previously published online at DOI: 10.1373/clinchem.2003.021469

[1] Nonstandard abbreviations: GUM, *Guide to the Expression of Uncertainty in Measurement*; PRL, prolactin; TSH, thyroid-stimulating hormone; MODUS, Model for Modular Evaluation of Uncertainty; EQA, external quality assessment; T_3, triiodothyronine; and T_4, thyroxine.

tient sample with a calibrator that has a known value of the measurand. Thus, the origin of the value assigned to the calibrator is very important for the result of measurement. Some calibrators are prepared in-house by weighing, dissolving, and diluting to a known volume, and in these cases the value can be calculated based on knowledge about the preparation procedure. This is typically an option for substances that are readily available in high purity, e.g., glucose. More often the value of the calibrator is assigned by a measurement, in other words, by another comparison. The comparisons may continue several steps, and the chain of comparisons can have several possible endings. For example, it may end in a certified reference material, such as the NIST SRM 1951a (lipids in human serum). The values assigned to SRM 1951a have been found by a definitive method, which means that they in effect have been compared with the definition of the corresponding SI unit (the mole). Another ending could be an International Standard prepared under the auspices of WHO. The WHO International Standards have values in arbitrary units that are established in a collaborative study. Often such high-level endings do not exist, however. For example, the value of a calibrator may be assigned based on the manufacturer's best measurement procedure.

These examples of "comparison chains" illustrate the property of metrologic traceability of the result of measurement *(3)*. Hence, the measurement carried out on the patient's sample is just the final one in a longer series of comparisons. Traceability is an important property of results. Almost 25 years ago, Tietz *(4)* pointed out that two results obtained by different measurement procedures at different times and different locations are comparable via their traceability to a common reference standard. In measurement, one should therefore strive for traceability to (globally) recognized standards, preferably the SI units. However, traceability alone is not enough to assure comparability. Each comparison made in the traceability chain causes uncertainty of the result. The accumulated uncertainty of the traceability chain must therefore be considered together with, or rather combined with, the uncertainty associated with the final measurement procedure.

Promises and Challenges from Traceability and Uncertainty

If measurement results are truly comparable through the means described above, there will be no need to repeat measurements when moving patients between healthcare centers, thereby reducing the inconvenience inflicted on the patients and reducing the amount of work required from the clinical laboratories. Moreover, results can—no matter the measurement procedure that has been used to obtain them—be compared with common limits and reference intervals, providing economic savings for the society. And because of the comparability of results from different measurement procedures, analytical problems such as those caused by interfering substances can be found simply by measuring the suspect sample with another measurement procedure. Any significant interference effect would cause a discrepancy between the two results that is not explained by their uncertainties.

These are some of the reasons that the idea of uncertainty should not be dismissed so lightly. But even if the concepts of metrologic traceability and uncertainty are accepted, there are difficulties in realizing them in practice. To benefit from traceability, for example, a calibration hierarchy has to be established for all clinically important measurands. Although the numbers of high-level calibrators and reference measurement procedures are growing steadily, there is still an overwhelming number of biochemical quantities that have no high-level ending of the traceability chain *(3)*. The establishment of these calibration hierarchies is a complex task, but fortunately not the primary task of routine clinical chemistry laboratories. However, when it comes to evaluation of uncertainty, the responsibility is clearly on the laboratory that produces the results. In the following, I will briefly review the GUM concept of uncertainty and in this connection respond to specific comments made in the accompanying commentary by Krouwer *(1)*. After that I will discuss two challenges that in my opinion are important to overcome. The first challenge is to make the uncertainty evaluation process more manageable, so that "routine" laboratories actually have the possibility in terms of capabilities and resources to estimate the uncertainty of the results of measurement; the second challenge is gaining access to the information about the uncertainty of calibrators supplied by external manufacturers.

Evaluation of (GUM) Uncertainty

Uncertainty (of a result of measurement) is defined in the GUM *(5)* as a parameter associated with the result of measurement, which characterizes the dispersion of values that can be reasonably attributed to the measurand. One should note from this definition that uncertainty is a property of the result of measurement, not a property of the measurement procedure. In fact, analytical imprecision is one of several components of the uncertainty. Note also that some familiar measures, for example, a SD or a confidence interval, both fulfill the criteria of the definition and that both can therefore be used to express uncertainty. However, to do arithmetic with uncertainties one needs to express the uncertainty as a SD, although the term used in the GUM is "standard uncertainty". Standard uncertainties are treated like standard deviations; in particular, their squares can be combined according to the mathematical rules for combining variances.

The principles of the uncertainty evaluation procedure are summarized schematically in Fig. 1, and the procedure has been described in detail elsewhere *(6–8)*. In the first step, significant sources of uncertainty (uncertainty components) are identified. Each uncertainty component is then assigned a standard uncertainty as defined above, and the standard uncertainties of all identified sources are

combined by classic "error-propagation" formulas to yield the standard uncertainty of the result of measurement. When reporting the uncertainty of the result, the combined standard uncertainty is multiplied with a so-called coverage factor, yielding an "expanded uncertainty". Usually a factor $k = 2$ is used because of the resemblance of the expanded uncertainty to a 95% confidence interval. Higher values of k can be chosen if a higher degree of coverage is wanted. Examples of uncertainty evaluations of relevance to laboratory medicine have been worked out for the measurement of glucose in blood *(9, 10)*, calcium in serum *(10)*, and prolactin (PRL) in serum *(11)*.

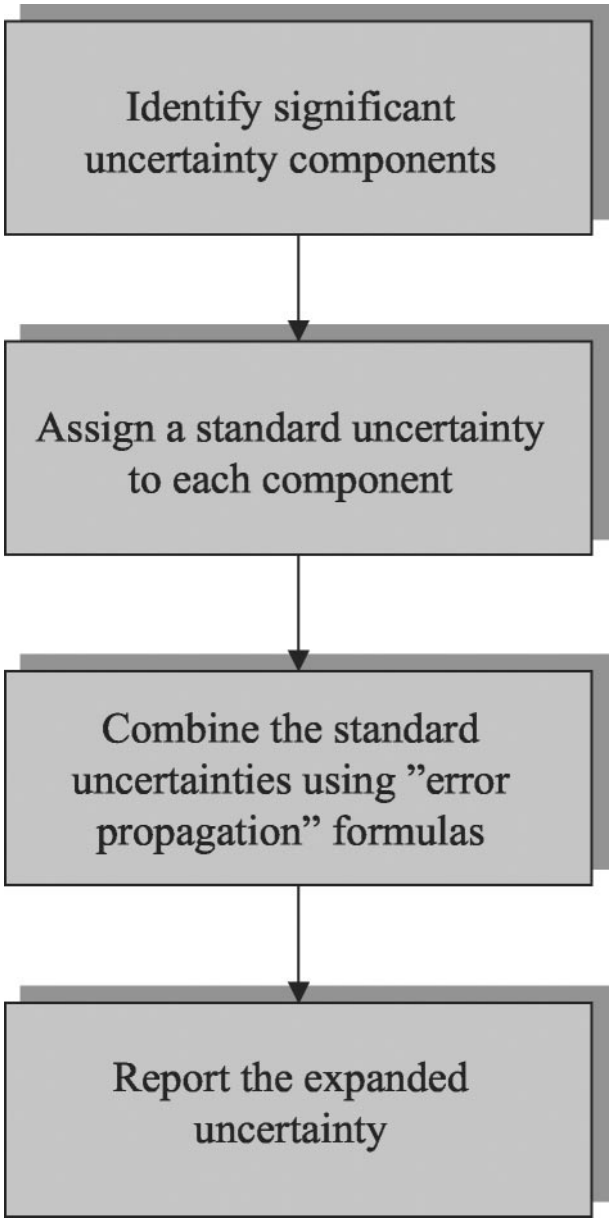

Fig. 1. Steps in the uncertainty evaluation process.

GUM Uncertainty and Bias

At this point some of the specific comments by Krouwer *(1)* merit comment. One of these comments regards systematic effects and the way they should be treated according to GUM. First, according to GUM, when a significant bias has been recognized and estimated, the result of measurement should be corrected by use of this estimate, and the uncertainty of the correction included in the uncertainty of the result *(5)*. Hence, GUM has only one way to treat recognized systematic effects and not three ways, as suggested by Krouwer *(1)*. Krouwer bases his suggestion on an example that I will briefly present here because it shows the importance of specifying the measurement procedure, including any corrections, when talking about uncertainty. The example is a sodium measurement method with the calibrator lot as a significant systematic effect. Krouwer mentions three ways that the laboratory may behave:

Laboratory 1 detects a significant bias in the calibrator lot and corrects the results.
Laboratory 2 uses the uncertainty statement on the certificate from the manufacturer to estimate the uncertainty of the calibrator value.
Laboratory 3 evaluates multiple calibrator lots in an experiment and uses an ANOVA model to calculate an uncertainty that can be assigned to the calibrator.

Apparently, Krouwer assumes that the systematic effect has been recognized in all three situations. In fact, however, only laboratory 1 has actually determined the bias. Hence, a correction is applied, and the uncertainty of the correction should therefore be included in the combined uncertainty of the result in accordance with the GUM approach.

In contrast, neither laboratory 2 nor laboratory 3 has estimated the bias (although the ANOVA data produced by laboratory 3 could be used in this task). They only have an uncertainty of the value of the calibrator. When making a measurement, both laboratories 2 and 3 must assume that the calibrator lot they are using is unbiased because this is the best estimate of the bias in probabilistic terms. Because they use the value of the calibrator in the measurement, the uncertainty of this value should be combined with the analytical imprecision to calculate the uncertainty of the result.

Thus, the evaluation of uncertainty depends on the actual way the result is produced. The consequences of this in terms of both the results and the accompanying uncertainty can be illustrated graphically by modeling the measurements made by laboratories 1 and 2 (laboratory 3 will produce results with a distribution similar to that of laboratory 2). The modeling parameters are presented in Table 1. The results of the modeling are shown in Fig. 2 in terms of the distribution of 5000 measurements on a sample with a sodium concentration of 150 mmol/L. Each measurement is done with a new calibrator lot. Laboratory 2 estimates the relative standard uncertainty of a

result by combining the relative standard uncertainty stated for the calibrator with the analytical imprecision:

$$\frac{u_{result}}{C_{result}} = \sqrt{\left(CV_A\right)^2 + \left(\frac{u_{cal}}{C_{cal}}\right)^2} = \sqrt{(5\%)^2 + (5\%)^2} = 7.1\%$$

As expected, the expanded ($k = 2$) uncertainty around 150 mmol/L includes ~95% of the distribution, and a larger uncertainty interval based on $k = 3$ includes almost all 5000 values (Fig. 2A). Hence, the estimated uncertainty is in excellent compliance with the metrologic definition of uncertainty presented above. In contrast, a 95% confidence interval based on the analytical imprecision includes <95% of the distribution (Fig. 2A). This means that if analytical imprecision is used to calculate a confidence interval around the result, then the "true" value of the measurand (the sodium concentration) will be outside the confidence limits more often than expected.

Next consider the measurement procedure of laboratory 1. The bias of the calibrator lot can be estimated by measuring each calibrator lot multiple time, e.g., 10 times, using a reference material with higher metrologic properties than the calibrator (see Table 1). The relative standard uncertainty of the bias estimate is thus:

$$\frac{u_{bias}}{f_{bias}} = \sqrt{\frac{(CV_A)^2}{n} + \left(\frac{u_{ref}}{C_{ref}}\right)^2} = \sqrt{\frac{(5\%)^2}{10} + (1\%)^2} = 1.87\%$$

Because laboratory 1 corrects for the bias, the uncertainty of the results is a combination of uncertainty of the bias estimate and the analytical imprecision, that is:

$$\frac{u_{result}}{C_{result}} = \sqrt{(CV_A)^2 + \left(\frac{u_{bias}}{f_{bias}}\right)^2} = \sqrt{(5\%)^2 + (1.87\%)^2} = 5.3\%$$

Because of the correction applied, the distribution of the results is narrower around the value of 150 mmol/L (Fig. 2B). However, the estimated standard uncertainty is reduced comparatively and still represents an excellent description of the values that are reasonable to attribute to the measurand.

These simple examples of uncertainty evaluations illustrate that talking about "calibrator lot bias" and other biases can cause a lot of confusion. What really matters for the results and their uncertainties is what you actually do (and do not do)! Do you estimate the bias or not? Whatever you choose, the examples show that the GUM approach ensures that a consistent uncertainty can be estimated.

Table 1. Parameters used in modeling measurements of sodium.[a]

Parameter	Symbol	Value
Relative analytical imprecision	CV_A	5%
Relative standard uncertainty of the value of the calibrator (the same for all lots)	$\frac{u_{cal}}{C_{cal}}$	5%
Relative standard uncertainty of the value of the reference material	$\frac{u_{ref}}{C_{ref}}$	1%

[a] The values were selected purely for illustrative reasons (see Fig. 2). They are not representative of real sodium measurements.

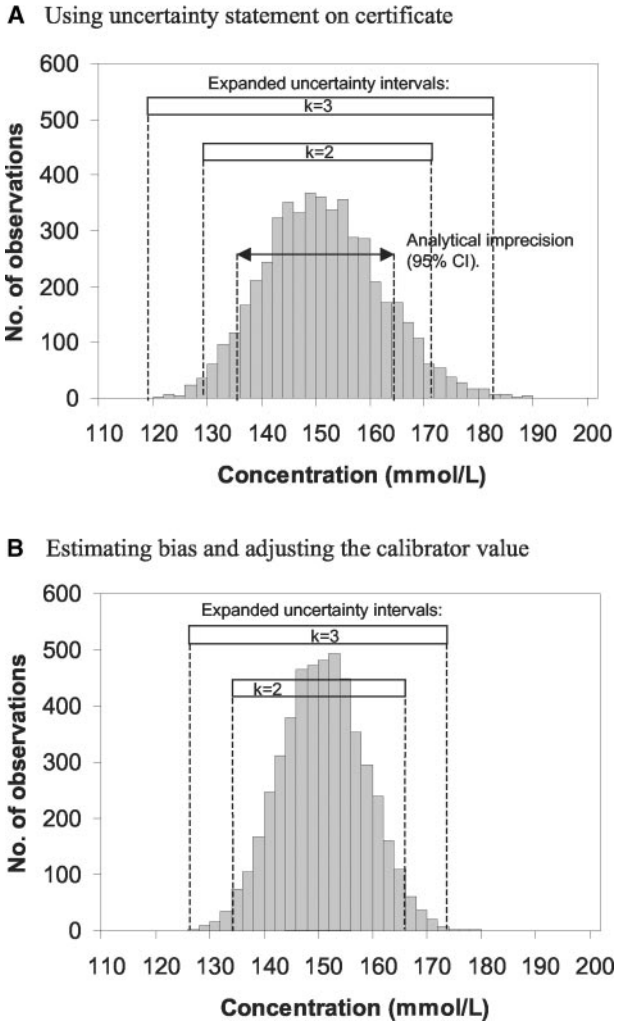

Fig. 2. Modeling of sodium measurements.

A sample with a sodium concentration of 150 mmol/L was measured 5000 times, each time with a different calibrator lot. The *histograms* show the distribution of individual results (for the parameters used, see Table 1). The *horizontal bars* represent expanded standard uncertainty intervals ($k = 2$ and 3). Calculation of the uncertainty was as follows: (*A*), no estimation of bias. The uncertainty of the calibrator lot was combined with the analytical imprecision to yield the uncertainty of the result. The *horizontal double arrow* indicates the 95% confidence interval calculated from the analytical imprecision. (*B*), estimation of bias. The uncertainty of the bias estimate was combined with CV_A in the calculation of the uncertainty of results. See the text for further details.

Outliers

Another issue brought up by Krouwer *(1)* concerns outliers. For a reason that is not clearly explained, Krouwer would like the GUM uncertainty interval to include outlying values, but this would run counter to the definition of uncertainty because an expanded uncertainty interval covering, e.g., 95% or even 99% of the values would obviously not include extreme values, which al-

most by definition are farther away from the "true value" than 3 SD. To be sure to include outliers one could, of course, further expand the uncertainty interval by use of $k \gg 3$, but the benefit of this is hard to see. One would ultimately want to detect outliers, not to include them in abnormally large uncertainty intervals. Outliers produced by diagnostic assays have important clinical implications because they may cause the wrong diagnosis to be made and lead to wrong treatment of the patient. Several studies have demonstrated that outlying results may occur relatively frequently. Ismail et al. *(12)*, for example, investigated sets of results from 5310 patients and found analytically incorrect results for thyroid-stimulating hormone (TSH) in 28 (0.53%) cases. The lack of specificity seems to be a common reason for outlying results, and immunoassays seem to be especially prone to interferences from substances present in the patient sample but not in the calibrators *(13)*. In the above-mentioned study by Ismail et al. *(12)*, interferences from unidentified substances were tested in three ways: (*a*) absence of parallelism on dilution with "analyte-free" sera; (*b*) changes in the result when analyzing the sample with heterophilic blocking agents added; or (*c*) differences between results when samples are analyzed by two different measurement procedures. In the latter test, a method-comparison study using patient samples without interference was used to compensate for systematic differences between the two measurement procedures. The three tests mentioned above are the standard "weapons" available to the laboratorian in the fight against interferences. However, with the patient waiting for a decision, one rarely has time to perform a method-comparison study. Checking suspect samples by use of a second measurement procedure may therefore be out of the question. However, when operating with results that are traceable to a common reference standard, and with appropriate GUM uncertainties worked out, such a checking procedure can actually work. Rapid detection of outliers is of clinical importance. The identification of measurement procedures that are susceptible to interfering substances will, in my opinion, compel manufacturers of these methods to work harder to improve them. Hence, uncertainty may in fact help to drive the improvement of the quality of such assays.

Standardization of the Uncertainty Evaluation Procedure

Although the principles of the uncertainty evaluation are easy to understand, the calculation may be difficult to carry out in practice. It takes time to evaluate uncertainty components. Moreover, the standard uncertainties are combined in different ways for different measurement methods, which adds to the complexity of the process. Krouwer *(1)* states that, ". . . ensuring that 'every effort has been made to identify such (systematic) effects' . . . is beyond the scope of most laboratories. . . ". Although I agree that evaluation of uncertainty can be demanding in terms of resources, I find Krouwer's interpretation of GUM too rigid. GUM per se does not preclude the evaluation of uncertainty of "routine" measurements. The efforts spent in evaluating the uncertainty should of course be reasonable and weighed against the purpose of the measurement and limitations in terms of time and economic resources. This is both common sense and also the usually accepted interpretation of the GUM approach in: for example, the international standard for accreditation of testing laboratories *(14)* and in the guide *Quantifying Uncertainty in Analytical Measurement (15)*. However, for the reasons mentioned, even allocating a "reasonable amount of time and effort" may not be enough. There is therefore much to gain if the uncertainty evaluation procedure can be standardized, including the systematic use of method validation data as input in the uncertainty evaluation procedure as proposed in *Quantifying Uncertainty in Analytical Measurement (15)*.

The use of validation data is systematized in the so-called Model for Modular Evaluation of Uncertainty (MODUS) method *(11)*. In brief, it recognizes that a clinical chemistry measurement in general consists of a measurement procedure applied to a sample and a calibrator. This generic model is illustrated in Fig. 3 and expressed mathematically in Eq. 1. For reasons explained in my original report *(11)*, the relationship between the value of the calibrator, sampling, and analysis is expressed as a multiplicative model:

$$C_{\text{result}} = C_{\text{analysis}} \cdot f_{\text{sampling}} \cdot f_{\text{traceability}} \cdot f_{\text{other}} \quad (1)$$

In this equation, C_{analysis} is the (usual) result of analysis, f_{sampling} is a correction factor for the bias introduced by the sampling process (including storage of the sample), $f_{\text{traceability}}$ corrects for the bias caused by systematic error in the value of the calibrator, and f_{other} is a correction factor that corrects for bias caused by other effects not encompassed by the previous term. When the sampling is unbiased, the value assigned to the calibrator is also unbiased, and other effects do not contribute to bias, then $f_{\text{sampling}} = f_{\text{traceability}} = f_{\text{other}} = 1$, or in other words, the outcome of the measurement (C_{result}) is the same as the result of the measurement procedure, C_{analysis}. However, this is not the case for their respective uncertainties. Assuming independence between the terms in Eq. 1, the corresponding "GUM expression" for the standard uncertainty of C_{result} (u_{result}) is conveniently expressed in terms of relative standard uncertainties:

$$\left(\frac{u_{\text{result}}}{C_{\text{result}}}\right)^2 = \left(\frac{u_{\text{analysis}}}{C_{\text{analysis}}}\right)^2 + \left(\frac{u_{\text{sampling}}}{f_{\text{sampling}}}\right)^2 + \left(\frac{u_{\text{traceability}}}{f_{\text{traceability}}}\right)^2 + \left(\frac{u_{\text{other}}}{f_{\text{other}}}\right)^2 \quad (2)$$

The relative standard uncertainty associated with analysis, $u_{\text{analysis}}/C_{\text{analysis}}$, can be estimated by the long-term relative analytical imprecision, CV_A. If not already known to the analyst, CV_A is easily accessible to experimental estimation. This uncertainty component integrates several

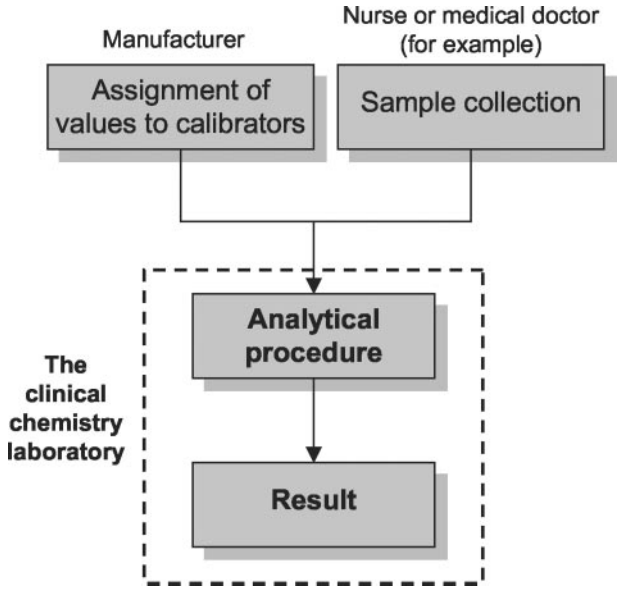

Fig. 3. Generic steps in a clinical chemistry measurement.
Adapted from Kristiansen *(11)*.

uncertainty components associated with the analysis (e.g., dilutions, injection, analysis, and estimation of the calibration function) and therefore saves time because there is no need to evaluate each of these components individually. The uncertainty of C_{result} also includes contributions from $f_{sampling}$ (the uncertainty of sampling), from $f_{traceability}$ (the uncertainty of the value assigned to the calibrator, which is an element in uncertainty associated with the traceability chain), and from f_{other}. For example, if the uncertainty from analytical drift is not included in the analytical imprecision, it should be included in the uncertainty of f_{other}.

The drawback of the simple model expressed by Eq. 1 is, of course, that there is no detailed information about the individual contributions from these components and there therefore is no basis for improving on the analysis. However, if the combined uncertainty is sufficient for the intended use of the method, this tradeoff is usually acceptable.

Analytical Assays from External Manufacturers

The presence of the factor $f_{traceability}$ in the MODUS model above emphasizes that traceability contributes uncertainty to the result of measurement. Information about the uncertainty of $f_{traceability}$ may be readily available if the calibrator is manufactured in-house, but it can be difficult to access if an external manufacturer makes the calibrator. Unfortunately, the calibrators in assays are typically delivered without a statement of uncertainty. Moreover, both the protocol for establishing traceability and the actual data are the property of the manufacturer and therefore usually not available to the laboratorian.

What should one do, then? The immediate solution is to refrain from estimating the uncertainty of the traceability chain. Because genuine comparability of results is based on traceability and an appropriate uncertainty statement, ignorance about this part of the measurement is not satisfactory. Therefore, in the long run, laboratories and responsible organizations should seek to exert influence on the manufacturers to make them disclose the necessary information. In Europe, the European Union has issued the EU Directive on In Vitro Diagnostic Devices (the IVD Directive) *(16)*, which assists laboratories and their organizations in this task. The directive obliges the manufacturers of in vitro diagnostic devices to ensure traceability to a reference at a higher level in the metrologic system. Implementation of the essential requirements of the directive is supported by several international standards, including a standard on metrologic traceability of calibrators *(17)*. Because it is impossible to establish traceability without assessing the uncertainty, it is likely that the IVD Directive will increase the focus on uncertainty and therefore will lead manufacturers to share this information with their customers more often than they do at present.

What are the consequences of the traceability chain lying "hidden" at the manufacturers? Stated qualitatively, the effects include systematic differences between results obtained by different measurement procedures (different assays). Quantitative estimates of these differences can be calculated from external quality-assurance (EQA) data, where results are grouped according to the measurement procedure. As discussed in Kristiansen *(11)*, it is likely that the uncertainty associated with traceability at the manufacturer would cause a proportional effect on the results of measurement; therefore, a multiplicative model should be appropriate:

$$C_{group\ i} = \mu \cdot \phi_{mf\ i} \cdot \epsilon_{group\ i} \quad (3)$$

where $C_{group\ i}$ is the average of results obtained by laboratories using the measurement procedure from the ith manufacturer, μ is the "true" value of the measurand, the factor $\phi_{mf\ i}$ is a constant factor that describes the relative bias associated with the measurement procedure of the ith manufacturer, and $\epsilon_{group\ i}$ is a random error caused by within- and between-laboratory variation among laboratories in the group. For the perfect assay, $\phi_{mf\ i}$ would be equal to 1. Statistical analysis of EQA data has confirmed that the proportional model given above was valid for measurements of PRL, TSH, triiodothyronine (T_3), and thyroxine (T_4) *(18)*. Fig. 4 shows the distribution of $\phi_{mf\ i}$ values for different measurement procedures for these four hormones.

An interesting measure is the ratio between the $\phi_{mf\ i}$ values of two different measurement procedures, because it corresponds to the mean ratio between results obtained by these procedures. This is sometimes referred to as "the bias between the two methods", which is misleading because usually neither of the two procedures is a reference measurement procedure. The ratio between all pairs of diagnostic procedures was calculated from the data in

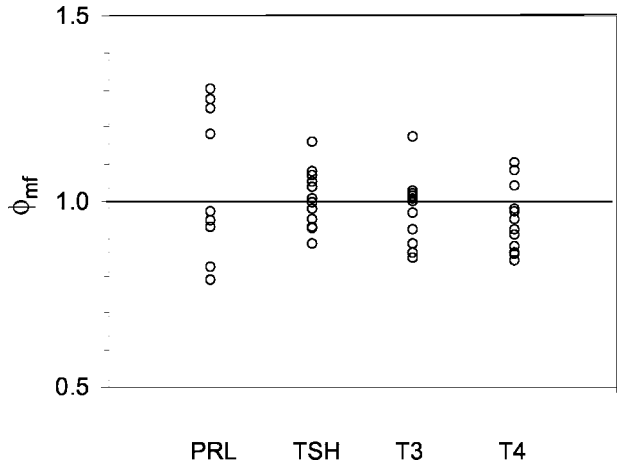

Fig. 4. Distribution of the manufacturer effect (ϕ_{mf}) for four hormone measurements.

Adapted from Kristiansen *(18)*.

Fig. 4 (only ratios >1 were considered), and values characterizing the distribution of these ratios are presented in Table 2. For PRL, for example, the relative difference between two different "average" measurement procedures is 27% (ratio between results is 1.27), whereas the interquartile range indicates that differences between 10% and 38% are typical. For the other hormones, the average relative difference is less, ~10–12%, and the corresponding interquartile range is also smaller. The results, however, indicate that relative differences up to ~30% are possible.

In comparison, Ismail et al. *(12)* reported an average ratio of 1.20 between TSH results obtained by the Abbot AxSYM and Bayer ACS-180, respectively. This systematic difference was found with patient samples, not EQA samples. This result is in the high end but still within the range of values found in Table 2.

The data in Fig. 4 and Table 2 do not indicate the nature of the observed systematic differences. PRL and TSH are both polypeptide hormones, whereas T_3 and T_4 are relatively low-molecular-weight structures based on tyrosine. In spite of the different natures of the molecules, there are no large differences in the distribution of the "manufacturer effects" (ϕ_{mf} values) between these hormones (Fig. 4). In comparison, different measurement procedures for cardiac troponin I may differ up to 100-fold *(19)*! As mentioned, the proportional nature of the differences observed in Fig. 4 is in accordance with the suggestion *(11)* that the underlying cause is the uncertainty of the manufacturer's traceability chain. Even the large differences observed for cardiac troponin I measurement procedures seem to be at least partly attributable to differences in calibration *(20)*. It therefore seems that there indeed are some valuable perspectives in implementing the IVD Directive *(16)*, thereby forcing manufacturers to focus more on traceability. It is no law of nature that different measurement procedures in clinical chemistry should differ systematically by 20%, 50%, or even more. A focus on traceability and stringent evaluation of uncertainty of measurement results may increase the pressure on manufacturers of measurement procedures to standardize their calibrators.

Table 2. Distribution of the ratio between results of measurements of two different assays.

Analyte	Average ratio (SD)	Interquartile range[a] (Q_1–Q_3)	Largest observation
PRL	1.27 (0.19)	1.10–1.38	1.65
TSH	1.10 (0.07)	1.04–1.13	1.31
T_3	1.12 (0.091)	1.04–1.17	1.38
T_4	1.12 (0.08)	1.05–1.16	1.31

[a] Range between the first (Q_1) and third (Q_3) quartiles.

Conclusions

The GUM approach to uncertainty deserves closer consideration. I have tried to emphasize some of the benefits that may come from the evaluation of uncertainty, but without ignoring the challenges. In my view, focusing on traceability and uncertainty has the potential to increase pressure on manufacturers of assays so that they increase their efforts to improve the quality of their products. This drive for improvement will include both the analytical quality, i.e., the specificity of the assays, and the metrologic quality of the calibrators. In my opinion, there are no fundamental problems in applying GUM uncertainty to measurements in laboratory medicine, and the practical problems, some of them discussed above, can be solved if there is a will. However, I recognize that there are many aspects to consider, and I hope that the debate on these matters will continue.

References

1. Krouwer JS. A critique of the GUM method of estimating and reporting uncertainty in diagnostic assays. Clin Chem 2003;49:1818–21.
2. International Organization for Standardization. International vocabulary of basic and general terms in metrology, 2nd ed. Geneva: ISO, 1993:60 pp.
3. Dybkaer R. From total allowable error via metrological traceability to uncertainty of measurement of the unbiased result. Accred Qual Assur 1999;4:401–5.
4. Tietz NW. A model for a comprehensive measurement system in clinical chemistry. Clin Chem 1979;25:833–9.
5. International Bureau of Weights and Measures, International Electrotechnical Commission, International Federation of Clinical Chemistry, International Organization for Standardization, International Union of Pure and Applied Chemistry, International Union of Pure and Applied Physics, International Organization of Legal Metrology. Guide to the expression of uncertainty in measurement, 1st ed. Geneva: ISO, 1995:101 pp.
6. Kallner A. Quality specifications based on the uncertainty of measurement. Scand J Clin Lab Invest 1999;59:513–6.
7. Dybkaer R. Setting quality specifications for the future with newer approaches to defining uncertainty in laboratory medicine. Scand J Clin Lab Invest 1999;59:579–84.
8. Kristiansen J, Christensen JM. Traceability and uncertainty in analytical measurements. Ann Clin Biochem 1998;35:371–9.

9. Kallner A, Waldenström J. Does the uncertainty of commonly performed glucose measurements allow identification of individuals at high risk for diabetes? Clin Chem Lab Med 1999;37:907–12.

10. Linko S, Örnemark U, Kessel R, Taylor PDP. Evaluations of uncertainty of measurement in routine clinical chemistry—applications to determinations of the substance concentration of calcium and glucose in serum. Clin Chem Lab Med 2002;40: 391–8.

11. Kristiansen J. Description of a generally applicable model for the evaluation of uncertainty of measurement in clinical chemistry. Clin Chem Lab Med 2001;39:920–31.

12. Ismail AAA, Walker PL, Barth JH, Lewandowski KC, Jones R, Burr WA. Wrong biochemistry results: two case reports and observational study in 5310 patients on potentially misleading thyroid-stimulating hormone and gonadotropin immunoassay results. Clin Chem 2002;48:2023–9.

13. Marks V. False-positive immunoassay results: a multicenter survey of erroneous immunoassay results from assays of 74 analytes in 10 donors from 66 laboratories in seven countries. Clin Chem 2002;48:2008–16.

14. International Organization for Standardization. General requirements for the competence of testing and calibration laboratories (17025). Geneva: ISO, 1999:26 pp.

15. Ellison SLR, Rosslein M, Williams A, eds. Quantifying uncertainty in analytical measurement. Eurachem/CITAC, 2001:119 pp.

16. Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on in vitro diagnostical medical devices. Off J Eur Communities 1998;L331:1–37.

17. International Organization for Standardization. In vitro diagnostic medical devices—measurement of quantities in biological samples—metrological traceability of values assigned to calibrators and control materials (17511). Geneva: ISO, 2003:23 pp.

18. Kristiansen J. Uncertainty is just repeated measurements. LabMedica Int 2002;19:7–10.

19. Panteghini M. Performance of today's cardiac troponin assays and tomorrow's [Editorial]. Clin Chem 2002;48:809–10.

20. Collinson PO, Boa FG, Gaze DC. Measurement of cardiac troponins. Ann Clin Biochem 2001;38:423–49.