Clinical Chemistry 64:3
455–464 (2018)

Special Reports

IFCC Working Group Recommendations for Assessing Commutability Part 2: Using the Difference in Bias between a Reference Material and Clinical Samples

Göran Nilsson,[1] Jeffrey R. Budd,[2] Neil Greenberg,[3] Vincent Delatour,[4] Robert Rej,[5] Mauro Panteghini,[6] Ferruccio Ceriotti,[7] Heinz Schimmel,[8] Cas Weykamp,[9] Thomas Keller,[10] Johanna E. Camara,[11] Chris Burns,[12] Hubert W. Vesper,[13] Finlay MacKenzie,[14] and W. Greg Miller,[15*] for the IFCC Working Group on Commutability

A process is described to assess the commutability of a reference material (RM) intended for use as a calibrator, trueness control, or external quality assessment sample based on the difference in bias between an RM and clinical samples (CSs) measured using 2 different measurement procedures (MPs). This difference in bias is compared with a criterion based on a medically relevant difference between an RM and CS results to make a conclusion regarding commutability. When more than 2 MPs are included, the commutability is assessed pairwise for all combinations of 2 MPs. This approach allows the same criterion to be used for all combinations of MPs included in the assessment. The assessment is based on an error model that allows estimation of various random and systematic sources of error, including those from sample-specific effects of interfering substances. An advantage of this approach is that the difference in bias between an RM and the average bias of CSs at the concentration (i.e., amount of substance present or quantity value) of the RM is determined and its uncertainty estimated. An RM is considered fit for purpose for those MPs for which commutability is demonstrated.

Background

Commutability was defined in part 1 of this series *(1)*. This second part describes a statistical procedure to assess commutability based on the difference in bias between a reference material (RM)[16] and clinical samples (CSs) measured using 2 different measurement procedures (MPs). This difference in bias is compared with a predefined criterion to make a commutability judgment. If more than 2 MPs are included in an assessment, the commutability is assessed pairwise for all combinations of 2 MPs. If 1 of the MPs is a reference measurement procedure (RMP), then commutability of the RM with each of the MPs can be assessed vs the RMP, and pairwise assessment among all MPs is not necessary.

As explained in part 1 of this series *(1)*, an MP refers to a written specification for how a measurement is performed. A measuring system is a physical in vitro diagnostic (IVD) medical device manufactured according to the MP specifications and used to make measurements on CSs. Results for an RM and for CSs measured using different measuring systems are used to assess commutability of an RM. For simplicity, in this series of reports we use the term MP when referring to either an MP or results from a specific measuring system that is an IVD medical device representative of the MP.

[1] Uppsala, Sweden; [2] Beckman Coulter, Chaska, MN; [3] Neil Greenberg Consulting, LLC, Rochester, NY; [4] Laboratoire national de métrologie et d'essais (LNE), Paris, France; [5] Wadsworth Center for Laboratories and Research, New York State Department of Health, and School of Public Health, State University of New York at Albany, Albany, NY; [6] Research Centre for Metrological Traceability in Laboratory Medicine (CIRME), University of Milan, Milan, Italy; [7] Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy; [8] European Commission, Joint Research Centre (JRC), Directorate F, Geel, Belgium; [9] Queen Beatrix Hospital, Winterswijk, the Netherlands; [10] ACOMED statistic, Leipzig, Germany; [11] National Institute of Standards and Technology, Gaithersburg, MD; [12] National Institute for Biological Standards and Control, A Centre of the MHRA, Hertfordshire, UK; [13] Centers for Disease Control and Prevention, Atlanta, GA; [14] Birmingham Quality/UK NEQAS, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK; [15] Department of Pathology, Virginia Commonwealth University, Richmond, VA.

* Address correspondence to this author at: P.O. Box 980286, Richmond, VA 23298-0286. Fax 804-828-0375; e-mail gmiller@vcu.edu.

Disclaimer: The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official views or positions of the Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry.

Received June 2, 2017; accepted December 15, 2017.
Previously published online at DOI: 10.1373/clinchem.2017.277541

[16] Nonstandard abbreviations: RM, reference material; CS, clinical sample; MP, measurement procedure; RMP, reference measurement procedure; IVD, in vitro diagnostic; C, commutability criterion; MSSD, mean square successive difference.

Currently applied approaches for commutability assessment use criteria based on the statistical distribution of results for CSs observed between pairs of MPs. Linear regression with prediction interval for the CSs has been commonly used to determine whether an RM is commutable *(2, 3)*. When an RM belongs to the same population as the CSs, an observation of the RM has a specified probability (usually set to 95%) to fall within the prediction interval limits. The prediction interval limits are dependent on the random errors in each MP and can be different for comparisons between different pairs of MPs, which prevents using a single criterion based on the intended use of the RM that is applicable to all MPs. Using a prediction interval for assessment of commutability is a test of the hypothesis that the RM can be considered to belong to the same population as the CSs. Not rejecting a hypothesis does not prove that it is true, and the prediction interval approach does not quantify how closely the RM agrees with the average relationship for the CSs at the concentration of interest. The uncertainty of the closeness of agreement is neglected in the prediction interval approach. An RM with a bias exactly equal to the prediction limit has a probability of approximately 50% to have a value within the prediction limits. For these reasons, a different approach is presented in this report.

An advantage of the approach described here is that the difference in bias between an RM and the average bias of CSs at the concentration of the RM is determined. This approach allows more relevant assessment of commutability being suitable for the intended use of an RM with the same criterion being used for all combinations of MPs included in the assessment. Another advantage is that the criterion to make a commutability judgment can be based on medically relevant differences between RM and CS results, which is not the case with the prediction interval approach.

In practice, an assessment of commutability cannot include all possible performance conditions for an MP, such as reagent lots, calibrator lots, and environmental conditions. We must restrict the assessment to measurements under specified conditions. In this report, the specified conditions are 1 run with each of the MPs using 1 lot of reagents and calibrators, and each MP operated and performing according to the specifications of its manufacturer. The conclusions about commutability of the RM are generalized to all future results using other IVD medical devices representative of the same MP with the assumption that other IVD medical devices have equivalent performance when operated under conditions such as different reagent and calibrator lots, maintenance, and operators. Limitations of this assumption were discussed in part 1 of this series *(1)*.

A worked example and explanation of the example calculations are provided in the Commutability Example Calculations and Commutability Example Explanation sections of the Data Supplement that accompanies the online version of this article at hiip://www.clinchem.org/content/vol64/issue3.

Models, Experimental Designs, and Assumptions

ASSESSMENT OF COMMUTABILITY

The experimental design considers the comparison of results, x and y, obtained by 2 MPs. In a typical experimental design, n CSs are measured in 1 run with each of the MPs. A simple model for the difference between single determinations of a CS is:

$$y - x = b(\mu) + d + e_y - e_x, \quad (1)$$

where:

μ is the true concentration of the CS

$b(\mu)$ is a common bias between the runs with the 2 MPs (the bias can be expressed by a continuous function of μ or a constant)

d is an error component specific for the CS (can be considered as a random component in a population of samples)

e_x is a within-run component of variation for MP x

e_y is a within-run component of variation for MP y.

The $b(\mu)$ term is the part of the difference that can be expressed by a continuous function of μ. Continuous means that small changes in μ result in small changes in $b(\mu)$. The term d is a sample-specific difference. The SD of component d is denoted σ_d. The terms e_y and e_x are the within-run components of variation with SDs $\sigma_{e(x)}$ and $\sigma_{e(y)}$, in the following denoted σ_x and σ_y. Sample-specific differences can be reduced only by an improved selectivity of 1 or both MPs.

Ideally, the variation within runs should be completely random, and the SDs σ_x and σ_y can be estimated from repeated measurements. Under this ideal situation, the SD of the sample-specific differences, σ_d, can be estimated according to an approach suggested by Nilsson *(4)*. If the prerequisite of random variation is not satisfied, position effects within measurement runs must be added to the model in Eq. 1. If the replicate measurements of the CSs are adjacent, the σ_d also includes the SDs of the position effects. Miller et al. *(5)* suggested to estimate the influence of possible position effects by performing a second run with the CSs in a randomized order in relation to the first run.

For commutability assessment, we recommend only 1 run and that the replicate measurements of the CSs are adjacent, i.e., made one after the other, and that position effects are investigated from measurements of the RMs made in different positions. The terms $b(\mu)$ and σ_d will usually depend on the specified population of CSs. The SDs are assumed to be at least approximately independent of the concentration. The distributions of the error components are assumed to be approximately normal.

COMMUTABILITY CRITERION

Commutability of the RM concerns how close the systematic difference (the bias) between the 2 MPs for the RM is to the average bias for the CSs, $b(\mu)$, at the concentration of the RM. The difference between the bias for the RM and the average bias for the CSs is denoted d_{RM} and expresses the closeness of agreement between the bias for the RM and the bias for the CSs.

For assessment of commutability, we need to specify a maximum value of $|d_{RM}|$ for the RM to be considered commutable. This maximum value is called the commutability criterion (C). For example, C can be the maximum acceptable bias when one intends to use the RM as a calibrator.

The SD of the contributions from the random components e and d to the differences between the 2 MPs, when the measured values are means of k replicates, is:

$$\sigma_{Random} = \sqrt{\sigma_d^2 + \frac{\sigma_x^2}{k} + \frac{\sigma_y^2}{k}} \quad (2)$$

Assuming a normal distribution, about 5% of the CS differences will be larger than $2\sigma_{Random}$ even if there is no bias between the measurement procedures for the CSs.

For assessment of commutability vs C, we need an estimate of d_{RM} (for simplicity, we use the same symbol for the estimate) and the expanded uncertainty $U(d_{RM})$ of the estimate. We will have 1 of the following 3 conditions:

1. The RM is commutable when the uncertainty interval $d_{RM} \pm U(d_{RM})$ is within $0 \pm C$.
2. The RM is noncommutable when the uncertainty interval $d_{RM} \pm U(d_{RM})$ is outside $0 \pm C$.
3. A commutability decision is inconclusive when the uncertainty interval $d_{RM} \pm U(d_{RM})$ and $0 \pm C$ are overlapping.

When $U(d_{RM}) > C$, it will not be possible to verify commutability. A too large $U(d_{RM})$ can be caused by (*a*) an unsuitable experimental design (too few replicates and/or number of CSs) or (*b*) poor precision and/or poor selectivity (large sample-specific differences). In the first case, the experimental design should be reconsidered and the experiment repeated with a new design. In the second case, 1 or both MPs should be excluded from the commutability assessment. The precision and selectivity of each MP should be known in advance and considered in the experimental design. As explained in part 1 of this series *(1)*, MPs should be prequalified for inclusion in a commutability assessment, and MPs with inadequate performance should be excluded. However, sample-specific differences are estimated in this commutability assessment experimental design and, when excessive, can be a reason to exclude an MP from the data analysis and to declare that the RM is not suitable for use with that MP.

EXPERIMENTAL DESIGN

The experiment should be designed in such a way that it is possible to obtain a reliable estimate of the value d_{RM} and the uncertainty of the estimate that is derived from an estimate of the bias for the RM and for the CSs and their uncertainties between 2 MPs. For identification of the main sources contributing to the differences between the MPs, it is also valuable to have estimates of the SDs for replicates, position effects, and sample-specific differences. To obtain these estimates, the following experimental design is recommended.

A number, n, of CSs are measured in k sequential adjacent replicates in 1 run with both MPs. The sequence of measuring the CSs must be randomly assigned regarding the concentrations (not in order of concentration) to avoid the possibility of confounding between the measuring order and a trend in the bias. Each RM is measured in p groups of k sequential adjacent replicates. The p groups shall be spread out in the run, i.e., the groups represent different positions in the series of measurements constituting a run. If there are known factors that may cause systematic differences between measurements, these factors must be considered in the design and statistical analysis of the experiment. For example, if a multichannel pipette is used for adding samples and/or reagents, it is suitable to have an equal number of replicates from each channel, i.e., k must at least be equal to the number of channels.

The number of CSs, n, must be large enough and the concentrations "evenly" distributed in the interval to make it possible to distinguish between sample-specific differences and a common bias between the MPs expressed by the continuous function $b(\mu)$. It is not necessary for the CSs to cover the measuring interval of the MPs, but it is important that the CSs cover a reasonable concentration interval around each RM to allow a reliable estimate of bias between the MPs. An essential prerequisite for the statistical analysis is that the change in the bias function between consecutive CS concentrations is relatively small compared with the variation around this function in a difference plot. Whether this prerequisite is satisfied can usually be judged from a visual inspection of the difference plot; see the Transformation of the Data section. The n CSs should usually be at least 30 to satisfy these requirements.

The minimum number of replicates, k, is 2, but triplicate measurements are recommended to allow removal of an outlier without removing all data for a CS or an RM position. If precision of MPs is a major uncertainty source, performing a larger number of replicates might be necessary.

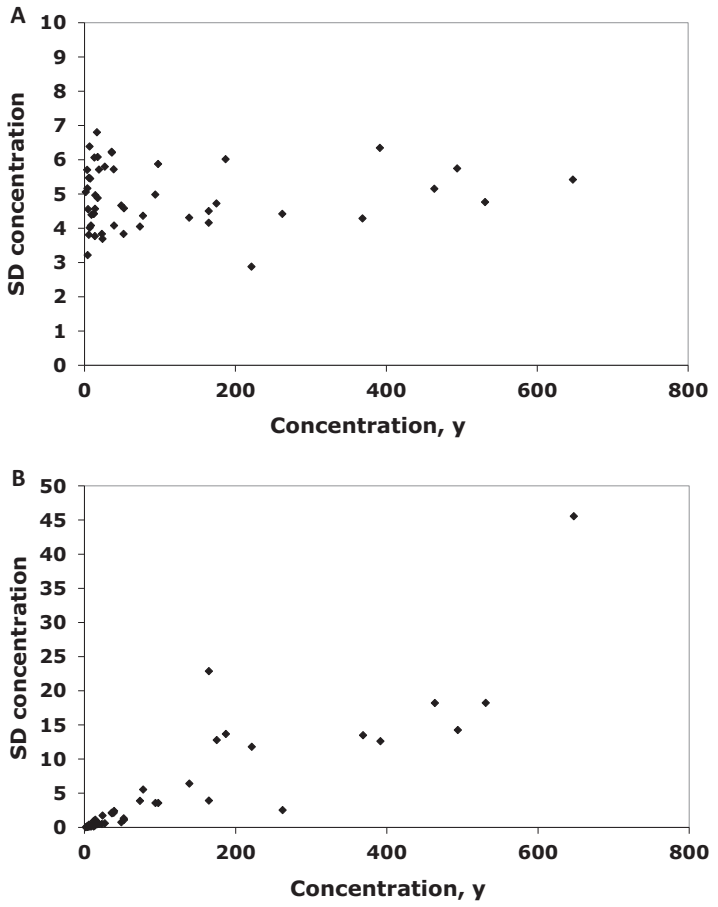

Fig. 1. Precision profiles as SD vs concentration.
(A) shows an approximately constant SD over the concentration interval. (B) shows a proportional relationship between SD and concentration.

Measuring the RMs in different positions makes it possible to investigate position effects and better estimate the uncertainty of the bias for an RM. The uncertainty of the bias estimate depends on the SD of the position effects and the number of degrees of freedom must be adequate for the estimate of this SD. We recommend that the number of positions, p, should be at least 5.

The estimate of the commutability measure, d_{RM}, must be determined with acceptable uncertainty. As a minimum requirement, we suggest that $U(d_{RM})$ should be $< C/2$. The suggested minimum values of n, k, and p do not guarantee that this requirement is satisfied, but the experimental design makes it possible to identify the dominating contribution to the uncertainty and indicates where an improvement of the experimental design is needed. An example of an allocation of measurements with $n = 50$, $k = 3$, $p = 5$, and 5 RMs is given as an example in the Commutability Example Calculations, worksheet Allocation, found in the online Data Supplement.

TRANSFORMATION OF THE DATA

The statistical analysis uses difference plots with or without transformation of the data for the statistical analysis. Difference plots are preferred to regression because possible trends in bias or sample-specific effects over the concentration interval are better identified.

For the statistical analysis, it is an advantage if the SDs of the random components are independent of the concentration. This requirement is often approximately satisfied for either concentration or ln(concentration). Other types of transformations can be used, but the advantage of using ln(concentration) is that it is easy to interpret the results. If SDs and differences of ln(concentration) are multiplied by 100, we obtain approximate values of CVs and relative differences in percent for concentration. The decision whether to use concentration or ln(concentration) for the statistical analysis is based on the experimental results. A precision profile as shown in Fig. 1 can be constructed for each MP by calculating the

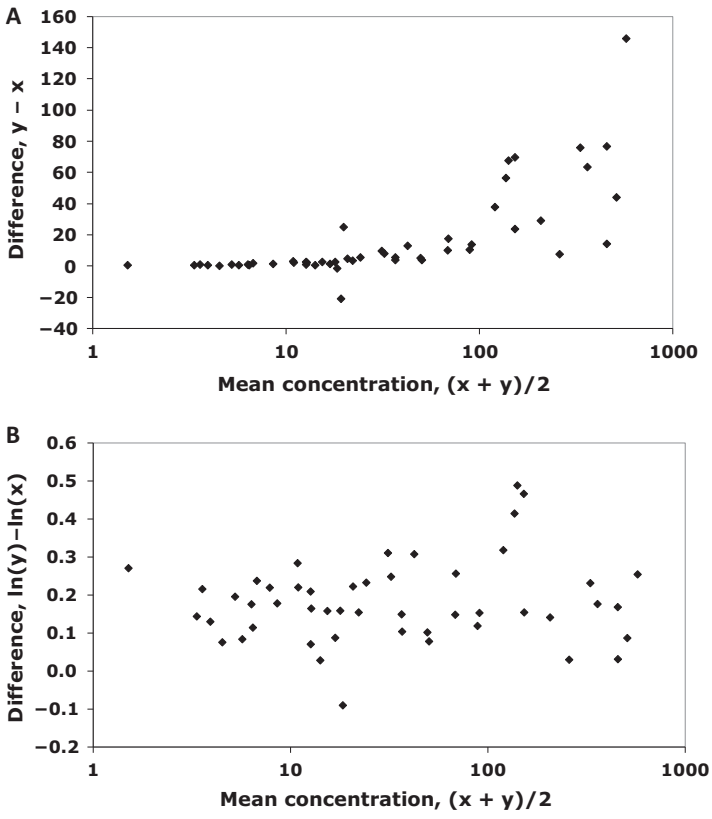

Fig. 2. Difference in concentration (A) and ln(concentration) (B) for the same data vs mean concentration of both MPs.
(A) is from spreadsheet tab CS__Trans found in the online Data Supplement. (B) is from spreadsheet tab CS__Trans (2) found in the online Data Supplement.

SD and the mean of the replicates for each CS and plotting the SDs against the means. If a precision profile does not indicate a strong (more than a factor 2) dependence between the SD and the concentrations, as in Fig. 1A, the SDs are pooled to a common estimate for each MP, denoted s_x and s_y, respectively. If the SDs seem to be proportional to the concentrations, as in Fig. 1B, it is an indication that ln(concentration) should be used; however, the next paragraph must also be considered.

Difference plots are examined for $y_i - x_i$ and $\ln(y_i) - \ln(x_i)$ on the y axis against x_i, when MP x is an RMP, or when neither of the MPs is an RMP, $(y_i + x_i)/2$ on the x axis. The width of the scatter is caused by the combined influence of the SDs of all random components. From a visual inspection of the 2 difference plots, one identifies the plot where the scatter width has the smallest dependence on the concentration to determine whether concentration or ln(concentration) should be used for the statistical analysis. Fig. 2 shows difference plots for $y_i - x_i$ and $\ln(y_i) - \ln(x_i)$ against the mean concentration. It is obvious in this case that ln(concentration) is preferred because of the consistent scatter over the concentration interval. It is more important that the width of the scatter is independent of the concentration for a difference plot than for the precision profiles in the previous paragraph. If the scatter width is independent of the concentration but precision profiles are not independent of the concentration, the contributions from the SD within triplicates, σ_x and σ_y, should be negligible. If neither concentration nor ln(concentration) gives an approximately constant scatter width, the statistical analysis should be performed for different concentration intervals, within which the scatter width is approximately constant. When transformation to ln(concentration) is used, the statistical analyses are performed with the ln(concentration) values. In precision profiles and difference plots, it is appropriate to have concentration on the x axis (possibly with a logarithmic scale) and use the ln(concentration) values on the y axis.

The decisions whether to transform and whether to perform the statistical analysis for different concentration intervals are based on subjective judgments. Models and

assumptions are always approximations of reality, and it is better to use a subjective judgment than to use a fixed model with no judgment at all.

HANDLING OF OUTLIERS

Outliers are observations that are distant from the main part of the observations. An outlier may be because of occasional problems in an MP, experimental mistakes, mix-up of results, transcription errors, or other types of operator errors. If such a cause is identified, the observation should be corrected or excluded. There may, however, be outliers for which the causes cannot be identified; for instance, when some of the CSs in a comparison of MPs have properties that cause ≥2 separate distributions of the differences between the MPs. To determine how close the bias for the RM is to the average bias for the CSs is not meaningful if the average bias represents 2 populations of CSs. The populations of CSs corresponding to these different distributions should, if possible, be identified (e.g., healthy and diseased donors) and commutability assessments performed for each population. If the outliers are relatively few (<10%), the only reasonable approach often is to exclude them. If the outlier results are not excluded, the calculation of the mean and the SD may be misleading.

Often a visual inspection is sufficient for identification of possible outliers. Obvious outliers can often be identified from, for instance, precision profiles and difference plots. If there are borderline cases, one can perform the analysis both without and with exclusions. If inclusion or exclusion of potential outliers gives essentially the same estimates, the observations can be included.

COMPONENTS OF VARIATION WITHIN RUNS ESTIMATED FROM THE RM

For each RM, we have p positions with k replicates in each position. First, the mean and SD for each position are calculated. The SD of the position means is denoted $s_{Pos\text{-}mean}$, and the pooled SD of the SDs within positions is denoted s_e. When the number of replicates is the same in each position, the pooled variance is the mean of the variances in the different positions. The pooled SD is the square root of the pooled variance. If there are no position effects, both $\sqrt{k} \cdot s_{Pos\text{-}mean}$ and s_e should be estimates of σ_e (the SD within positions), and to test the hypothesis of no position effects, the test statistic is:

$$F = \frac{k s^2_{Pos\text{-}mean}}{s^2_e} \tag{3}$$

The SD of the position effects is estimated by:

$$s_{Pos} = \sqrt{s^2_{Pos\text{-}mean} - \frac{s^2_e}{k}} \tag{4}$$

When the value under the root sign is negative, ($F < 1$) s_{Pos} is set to 0. When >1 RM is included, pooled estimates can be calculated as the square root of the mean variances. When calculating the pooled SD of the position effects, a negative value of the variance under the root sign in Eq. 4 shall not be replaced by 0.

ESTIMATE DIFFERENCES BETWEEN MPs FROM THE CSs

The analysis is performed for either concentration or ln(concentration). In the following, x_i and y_i denote either the mean of concentration or ln(concentration) for sample i. From the replicates of each CS, an estimate of SD between replicates is obtained. A pooled estimate of the SD between replicates is obtained from all CSs by calculating the square root of the mean of the variances for the individual CSs. The pooled SDs between replicates are denoted s_x and s_y, respectively. If the SDs between replicates for the RM(s) do not differ significantly (F-test) from the pooled estimates for the CSs, then the SDs for the RM(s) can be included in s_x and s_y. Otherwise, separate estimates are used for the RM(s) and the CSs.

The differences $B_i = y_i - x_i$ are plotted against the mean concentrations of the 2 MPs or against x_i, when x represents an RMP. A difference plot can be characterized by 2 components: a continuous function fitted to the center of the scatter and the variation around the function, expressed by the SD. The first component is an estimate of $b(\mu)$. The SD of the scatter should be σ_{Random}, defined in Eq. 2.

It is often reasonable to assume that the maximum difference between consecutive concentrations of the CSs corresponds to a maximum change of the bias function, which is small compared with the contribution from the within-run variation to the observed differences B_i. Consecutive B values should then have approximately the same expected mean, and we can estimate the SD around the bias function from each pair of consecutive values. Pooling the estimates from all the consecutive differences gives the SD estimate based on the mean square successive difference (MSSD).

$$s_{MSSD} = \sqrt{\frac{1}{2(n-1)} \sum_{i=1}^{n-1} (B_{i+1} - B_i)^2}, \tag{5}$$

where B_i is ordered according to ascending values of x_i or $(x_i + y_i)/2$. s_{MSSD} is used as an estimate of σ_{Random}, which includes the variability associated with sample-specific differences and does not include the variability associated with a trend in differences. This estimate should be of the same size as the usual estimate of the SD of the B values, calculated as:

$$s_B = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (B_i - B_{CS})^2}, \tag{6}$$

when the difference between the measurement procedures is a constant (i.e., no trend). B_{CS} is the mean of the differences B_i. Thus, when s_{MSSD} is significantly smaller than s_B, a trend is indicated. For $n > 20$, the distribution of $(s_{MSSD}/s_B)^2$ is approximately normal with mean 1 and $SD = \sqrt{\frac{1}{n+1}\left(1 - \frac{1}{n-1}\right)}$; see Hald *(6)*. This test supports the decision regardless of whether the bias function is constant in a concentration interval.

When there are sample-specific effects, the s_{MSSD} is an estimate of the SD of the contributions from the random components *e* and *d* to the differences between the 2 MPs according to Eq. 2. When there are no sample-specific differences, s_{MSSD} should be an estimate of:

$$\sigma_{Random(0)} = \sqrt{\frac{\sigma_x^2 + \sigma_y^2}{k}} \quad (7)$$

and the hypothesis of no sample-specific differences can be tested by:

$$F = \frac{k \cdot s_{MSSD}^2}{s_x^2 + s_y^2} \quad (8)$$

where F has an F distribution.

With the suggested experimental design (triplicate measurements), k is equal to 3. The appropriate numbers of degrees of freedom in the numerator and the denominator are not obvious. s_{MSSD} is estimated from $n - 1$ differences, but they are not independent and the degrees of freedom must be less than $n - 1$. If s_{MSSD} is calculated from the differences between $B_2 - B_1$, $B_4 - B_3$, and so on, it is based on $n/2$ independent differences when n is even and $(n - 1)/2$ when n is odd. Thus, the number of degrees of freedom for the estimate s_{MSSD} should at least be equal to the integer part of $n/2$, and this value is used in the test.

In the denominator in Eq. 8, we have the sum of 2 variance estimates, and the Welch–Satterthwaite formula can be used to calculate the effective number of degrees of freedom *(7)*. In this application, the minimum number of the degrees of freedom for s_x and s_y is used. By using minimum numbers of degrees of freedom both in the numerator and the denominator, the F-test should be conservative (i.e., the probability of incorrectly rejecting the hypothesis of no sample-specific differences is less than the nominal significance level).

An estimate of σ_d, the SD of the sample-specific differences, is obtained by:

$$s_d = \sqrt{s_{MSSD}^2 - \frac{s_x^2 + s_y^2}{k}}. \quad (9)$$

If the expression under the root sign is negative, the estimate $s_d = 0$ is used.

When there are no trends, s_B may be used instead of s_{MSSD} in Eq. 9.

If there are position effects, these effects are included in the estimate s_d. By using the estimates of position effects for RMs, we can correct for these effects and obtain a corrected value [denoted $s_{d(corr)}$]

$$s_{d(corr)} = \sqrt{s_{MSSD}^2 - \frac{s_x^2 + s_y^2}{k} - s_{Pos(x)}^2 - s_{Pos(y)}^2}. \quad (10)$$

The commutability assessment and the calculation of the expanded uncertainty $U(d_{RM})$ are performed in the same way regardless of whether there are significant sample-specific differences. However, if the uncertainty is too large for a conclusive decision and the commutability assessment must be repeated, it is essential to identify the error component that gives the dominating contribution to the uncertainty. In other words, shall we increase the number of replicates or the number of CS or should we adjust the qualifications to include CS or exclude an MP from commutability assessment because of nonselectivity for the measurand. One of the advantages with this approach is that no assumption of a specific model for $b(\mu)$ is required for the estimation of the sample-specific differences. We can separate sample-specific differences from a common bias assumed to be a continuous function of the concentration.

COMMUTABILITY OF THE RM

The commutability of the RM is assessed as the difference between the bias for the RM and the average bias for the CSs, $b(\mu)$, between 2 MPs at the concentration of the RM. This difference is denoted d_{RM}. To obtain an estimate of d_{RM}, we need an estimate of the bias for the RM and $b(\mu)$ for the CSs. The bias for the RM is estimated by the observed difference $B_{RM} = y_{RM} - x_{RM}$, and the standard uncertainty is estimated by

$$u(B_{RM}) = \sqrt{\frac{s_{Pos\text{-}mean(x)}^2 + s_{Pos\text{-}mean(y)}^2}{p}}. \quad (11)$$

$s_{Pos\text{-}mean(x)}$ and $s_{Pos\text{-}mean(y)}$ are the SDs between position means for the RMs defined in the Components of Variation Within Runs Estimated from the RMs section.

The appropriate estimate of $b(\mu)$ at the concentration of the RM depends on the outcome of the experiment. The following outcomes are possible (note that the uncertainty of B_{RM} is the same):

A: The bias function $b(\mu)$ is approximately constant in the whole concentration interval, and the CSs bracket the concentration of the RM

$b(\mu)$ is estimated by B_{CS} and the standard uncertainty by:

$$u(B_{CS}) = \frac{s_B}{\sqrt{n}}. \quad (12)$$

d_{RM} is estimated by $B_{RM} - B_{CS}$ with the standard uncertainty:

$$u(d_{RM}) = \sqrt{\frac{s^2_{Pos\text{-}mean(x)} + s^2_{Pos\text{-}mean(y)}}{p} + \frac{s^2_B}{n}} \quad (13)$$

B: The bias function $b(\mu)$ is approximately constant in a concentration interval enclosing the RM

It is possible to find a concentration interval with q CSs bracketing the RM where $b(\mu)$ seems to be approximately constant. In this case, q must be large enough to give an acceptable uncertainty. We have the same situation as in A but with q CSs instead of n; thus, we have:

$$u(d_{RM}) = \sqrt{\frac{s^2_{Pos\text{-}mean(x)} + s^2_{Pos\text{-}mean(y)}}{p} + \frac{s^2_B}{q}} \quad (14)$$

C: The bias function $b(\mu)$ has an approximately linear trend in an interval where there are $q/2$ CSs on each side of the RM

The number q is even and must be large enough to give an acceptable uncertainty. The CSs should also be "evenly" distributed in the concentration interval.

The mean bias of the q CSs bracketing the RM is used as an approximate estimate of $b(\mu)$ at the concentration of the RM, but if the uncertainty is calculated according to Eq. 12, it will be an overestimation, as the trend in $b(\mu)$ contributes to s_B. Instead of s_B, it is more reasonable to use s_{MSSD} according to Eq. 5, and the uncertainty is given by:

$$u(d_{RM}) = \sqrt{\frac{s^2_{Pos\text{-}mean(x)} + s^2_{Pos\text{-}mean(y)}}{p} + \frac{s^2_{MSSD}}{q}} \quad (15)$$

A large magnitude in the bias trend indicates a severe problem, and the commutability assessment may not be possible.

D: The prerequisites for situations A, B, and C are not satisfied (e.g., we have none or only a few CSs with concentrations close to that of the RM)

It may be tempting to fit a model to $b(\mu)$ and extrapolate or interpolate to the concentration interval of interest. Forcing the data into a model is not acceptable. With no or few CSs in the relevant concentration interval, we have no possibilities to verify that the model is reasonable.

DETERMINING THE COVERAGE FACTOR FOR THE UNCERTAINTY ESTIMATE

To evaluate commutability, we need the expanded uncertainty $U(d_{RM})$ obtained by multiplying $u(d_{RM})$ by a suitable coverage factor. If we want a risk of 5% to erroneously classify an RM just outside C as commutable, the coverage of the expanded uncertainty shall be 90%. An uncertainty interval $d_{RM} \pm U(d_{RM})$ within $0 \pm C$ should be equivalent to rejecting the hypothesis that $|d_{RM}| > C$ at the 5% level of significance. The standard uncertainties in Eqs. 13, 14, and 15 are a combination of 3 components, which makes the calculation of the coverage

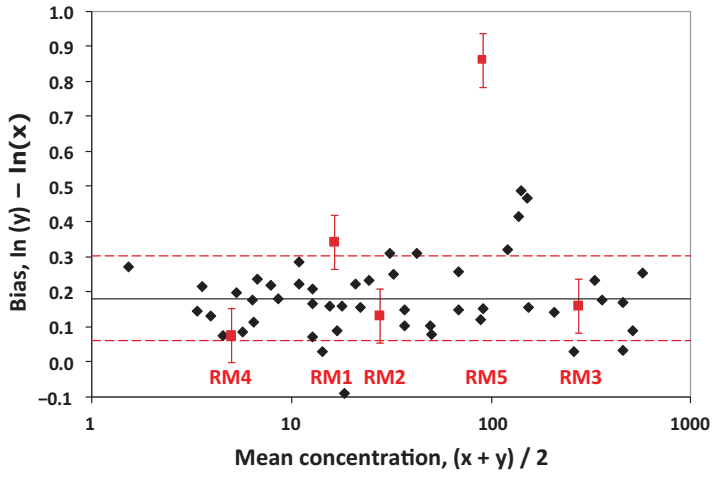

Fig. 3. Difference in bias between RMs (red squares) and CSs (black diamonds) vs mean concentration of the 2 measuring systems.
The solid black line is the mean bias between the 2 measurement procedures for the CSs. The red dashed lines are the commutability criteria. The red squares are the mean bias between the 2 MPs for the RMs, and the bars are the uncertainty in the difference in bias between RM and CS mean bias. RM1, RM2, and RM4 are indeterminate; RM3 is commutable; RM5 is noncommutable. Fig. 3 is from the spreadsheet tab CS&RM__Diff found in the online Data Supplement.

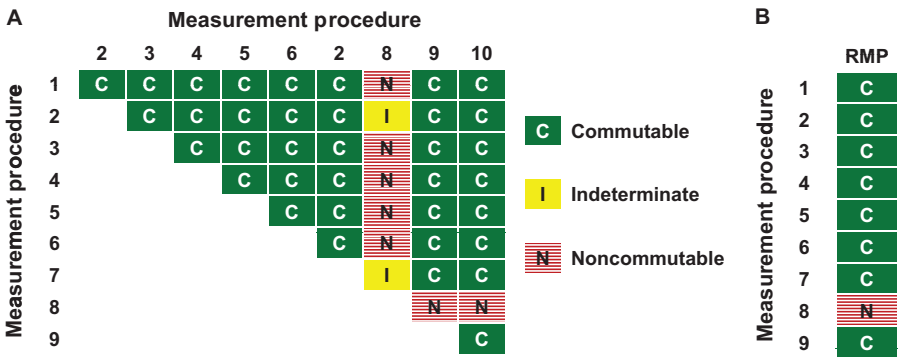

Fig. 4. Summary of commutability conclusions for a RM for multiple MPs.

(A) shows a representation of commutability conclusions for all combinations of pairs of MPs. (B) shows a representation when an RMP is available. The notation C, I, N could be replaced with numeric values for difference in bias and its uncertainty if desired.

factor complicated. However, with $p \geq 4$ (pooled estimates of $s_{Pos\text{-}mean}$ from at least 2 RMs) and n (or q) ≥ 12, a coverage factor of 1.9 gives a coverage of at least 90%. This coverage factor is suggested when there are no reasons for a larger value. With an infinite number of degrees of freedom, the coverage factor is about 1.7. The sizes of n, q, and p discussed in this section concern the minimum values, which can justify a coverage factor of 1.9. The numbers recommended in the experimental design section meet these minimum sizes.

Presentation of Commutability Assessment

A suitable way to illustrate the results from a commutability assessment is shown in Fig. 3 based on the example data provided in the online Data Supplement. The mean bias between the 2 MPs for the CSs is shown as a solid line (black). The dashed lines (red) are the commutability criteria. The squares (red) are the mean bias between the 2 MPs for the RMs, and the bars represent the uncertainty in the difference in bias between RM and CS mean biases. RM1, RM2, and RM4 are indeterminate; RM3 is commutable; RM5 is noncommutable.

Commutability Assessment for More Than Two Measurement Procedures

In most cases, more than 2 MPs are included in a commutability assessment, and all combinations of 2 MPs are evaluated as pairs as described here and shown in Fig. 3 for 1 pair. The commutability of an RM determined for all combinations of MPs examined can be presented in different ways. One convenient summary representation for all combinations of pairs of MPs is shown in Fig. 4A. In this example, the RM is not commutable for MP8 and is commutable for the other MPs in the assessment. In cases when an RMP is available for a measurand, all clinical laboratory MPs can be compared with the RMP (see Fig. 4B), and it is not necessary to compare all combinations of MP pairs with each other. Considerations for the fraction of MPs for which an RM should be commutable to be suitable for its intended use were discussed in part 1 of this series.

Conclusion

The approach to commutability assessment described here gives estimates of the differences in bias between RMs and CSs when 2 MPs are compared. The uncertainties of these estimates are also calculated. A single fixed criterion for commutability of an RM can be applied to all combinations of pairs of MPs. The criterion can be selected based on the intended use of an RM as a calibrator, trueness control, or external quality assessment material, and the criterion can be related to the requirements for medical decisions based on the laboratory test results. The commutability assessment determines whether the difference in bias plus its uncertainty fulfills the criterion for a conclusion that an RM is commutable, noncommutable, or indeterminate for pairs of MPs. Conclusions regarding suitability for use of an RM can be made by assessing its commutability for all MPs in the assessment as described in part 1 of this series *(1)*.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: R. Rej, *Clinical Chemistry*, AACC; C.J. Burns, National Institute for Biological Standards and Control; W.G. Miller, *Clinical Chemistry*, AACC.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: None declared.
Expert Testimony: None declared.
Patents: None declared.
Other Remuneration: N. Greenberg, College of American Pathologists.

Role of Sponsor: No sponsor was declared.

References

1. Miller WG, Schimmel H, Rej R, Greenberg N, Ceriotti F, Burns C, et al. IFCC working group recommendations for assessing commutability part 1: general experimental design. Clin Chem 2018;64:447–54.
2. Evaluation of commutability of processed samples; approved guideline. 3rd Ed. CLSI document EP14-A3. Wayne (PA): Clinical and Laboratory Standards Institute; 2014.
3. Characterization and qualification of commutable reference materials for laboratory medicine; approved guideline. CLSI document EP30-A. Wayne (PA): Clinical and Laboratory Standards Institute; 2010.
4. Nilsson G. Comparison of measurement methods based on a model for the error structure. J Chemometrics 1991; 5:523–36.
5. Miller WG, Thienpont LM, Van Uytfanghe K, Clark PM, Lindstedt P, Nilsson G, Steffes MW. Toward standardization of insulin immunoassays. Clin Chem 2009;55: 1011–8.
6. Hald A. Statistical theory with engineering applications. New York (NY)/London (UK): John Wiley & Sons; 1952.
7. Evaluation of measurement data– guide to the expression of uncertainty in measurement. Joint Committee for Guides in Measurement (JCGM). 2008. Sevres, France: International Bureau of Weights and Measures.