IFCC Working Group Recommendations for Assessing Commutability Part 3: Using the Calibration Effectiveness of a Reference Material

Jeffrey R. Budd,[1] Cas Weykamp,[2] Robert Rej,[3] Finlay MacKenzie,[4] Ferruccio Ceriotti,[5] Neil Greenberg,[6] Johanna E. Camara,[7] Heinz Schimmel,[8] Hubert W. Vesper,[9] Thomas Keller,[10] Vincent Delatour,[11] Mauro Panteghini,[12] Chris Burns,[13] and W. Greg Miller,[14*] 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 based on its ability to fulfill its intended use in a calibration traceability scheme to produce equivalent clinical sample (CS) results among different measurement procedures (MPs) for the same measurand. Three sources of systematic error are elucidated in the context of creating the calibration model for translating MP signals to measurand amounts: calibration fit, calibrator level trueness, and commutability. An example set of 40 CS results from 7 MPs is used to illustrate estimation of bias and variability for each MP. The candidate RM is then used to recalibrate each MP, and its effectiveness in reducing the systematic error among the MPs within an acceptable level of equivalence based on medical requirements confirms its commutability for those MPs. The RM is declared noncommutable for MPs for which, after recalibration, the CS results do not agree with those from other MPs. When a lack of agreement is found, other potential causes, including lack of calibration fit, should be investigated before concluding the RM is noncommutable. The RM is considered fit for purpose for those MPs where commutability is demonstrated.

Background

The goal of providing a higher order reference material (RM)[15] is to ensure that any clinical sample (CS) will have equivalent results across measurement procedures (MPs), within an uncertainty consistent with medical decision requirements. When the results for multiple CSs obtained with multiple MPs are compared, the differences are caused by the following types of errors:

- Random errors within MPs
- Sample-specific differences between MPs
- A bias between MPs (a function of the concentration)

The causes of the bias error can be an inappropriate model for the calibration curve, incorrect values assigned to the calibrators, and a difference in behavior between calibrators and CSs (not the same relationship between concentration and response). The bias error can be reduced by recalibration with a suitable RM or set of RMs. Consequently, the commutability of an RM can be assessed by how well bias among measurements of CSs is reduced when that RM is used in the calibration traceability schemes of different MPs. The approach to commutability assessment described here is applicable for RMs intended for use as calibrators in a calibration hierarchy described in ISO 17511 *(1)*.

The success of such an assessment process depends on how well each MP is designed and implemented. Therefore, the reasons for any bias remaining after reca-

[1] Beckman Coulter, Chaska, MN; [2] Queen Beatrix Hospital, Winterswijk, the Netherlands; [3] 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; [4] Birmingham Quality/UK NEQAS, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK; [5] Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy; [6] Neil Greenberg Consulting, LLC, Rochester, NY; [7] National Institute of Standards and Technology, Gaithersburg, MD; [8] European Commission, Joint Research Centre (JRC), Directorate F, Geel, Belgium; [9] Centers for Disease Control and Prevention, Atlanta, GA; [10] ACOMED statistic, Leipzig, Germany; [11] Laboratoire national de métrologie et d'essais (LNE), Paris, France; [12] Research Centre for Metrological Traceability in Laboratory Medicine (CIRME), University of Milan, Milan, Italy; [13] National Institute for Biological Standards and Control, A Centre of the MHRA, Hertfordshire, UK; [14] 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 18, 2017.
Previously published online at DOI: 10.1373/clinchem.2017.277558

[15] Nonstandard abbreviations: RM, reference material; CS, clinical sample; MP, measurement procedure; IVD, in-vitro diagnostic; IMPBR, intermeasurement procedure bias range; ID, identifier; $W_{(3)}$, Quasi range statistic.

libration with an RM should be investigated before the RM is characterized as noncommutable with an MP. For example, if the remaining bias is caused by nonselectivity of an MP, or by an inadequate design or implementation of the MP, then the measurement issues of that MP need to be addressed rather than concluding the RM is not suitable for use. To assess the commutability of an RM, estimates of systematic errors (bias) must be determined. This determination must be done within data sets where random variation is present and when sample interferences for specific samples on specific MPs may be seen. Therefore, throughout the following assessment, estimates of systematic error are determined using order statistics that are little affected by outlying results (e.g., median).

As explained in part 1 of this series *(2)*, 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.

Calibration Process and Sources of Error

Before clinically relevant results can be reported, the MP that produces those results must be calibrated with a suitable number of calibrator levels. Although CS-specific issues such as interferences are not relevant for the calibration process, random variation does play a role. The manufacturer must create a process that ensures true results by choosing a reproducible calibration scheme traceable to the highest order reference available. In the simplest case, a calibration relationship is established by measuring the signal created by the MP when calibrator samples of stated concentration (i.e., amount of substance present or quantity value) are tested and fit to a linear regression:

$$y = \alpha + \beta \cdot x_s + \varepsilon \tag{1}$$

where y is the signal, ε is the random variation in the signal, x_s is the stated concentration of the calibrator levels used, α is the linear regression intercept, and β is the linear regression slope. A more complex, 4-parameter logistic curve model is presented in the Data Supplement that accompanies the online version of this article at hiip://www.clinchem.org/content/vol64/issue3. Regardless of the type of calibration fit, the signal random variation ε can be characterized with a precision profile across the concentration interval, typically assuming a normal distribution on the signal axis.

During the calibration process, sources of systematic error include the following:

- The lack of fit of the mathematical model characterizing the relationship between signal and concentration. This lack of fit can be described by a function *f*, which is the ratio of the observed signal to the model fit. When the MP calibration model fits through all the signal results from each of the calibrator samples, then $f = 1$ across the concentration interval.
- The degree to which the stated concentrations of the calibrator samples are not true values, as defined by traceability to a specified RM. The ratio of stated to true concentration across the concentration interval can be described by a function *g*. When all MP calibrator sample values are assigned via a traceability scheme created with the specified RM, then $g = 1$ across the concentration interval. Such a value assignment is most accurate if $f = 1$. This report describes an approach to assessing the commutability of this RM when it is used in the calibration hierarchy of the calibrators used in a clinical laboratory MP.
- The extent to which the stated concentrations of the calibrator samples, traceable to the RM being assessed, are equal to measurement results from authentic CS having the same concentration. The extent of this equivalence, or commutability, can be described by a function *h*, which is the ratio of observed CS results vs their consensus target results. If, after setting $g = 1$ using the RM, CS samples give their consensus target results on a specified MP, then $h = 1$ across the concentration interval for that MP.

When the functions *g* and *h* are not equal to 1, then they are systematic error terms that modify the true concentration *x* such that its estimated value is

$$\hat{x} = x \cdot g \cdot h \tag{2}$$

The function *f* is an error term that characterizes the model fit to the signal response over the entire concentration interval. Although this function may change somewhat at the creation of each calibration curve, the goal of using this function is to describe a systematic calibration fit error that is integral to MP design. Because none of the error terms can be assumed to be constant across the measuring interval, the amount of error introduced by each error function depends on the position (*i*) on the calibration curve of signal vs concentration. This error at each concentration is described in the expanded equation

$$y_i \cdot f_i = \alpha + \beta \cdot x_i \cdot g_i \cdot h_i + \varepsilon_i \tag{3}$$

Curve fitting algorithms can use a closed-form solution such as a least-squares regression, a numerical solution such as a nonlinear iterative fit to a 4-parameter logistic curve, or a piecewise solution that can incorporate both approaches like a cubic spline fit. All such algorithms attempt to make f_i as close to unity as possible, typically using the signal precision profile ε_i to weight the fitting process. A common approach is to weight the fit to a calibration point (x_i, y_i) by the inverse of ε_i^2 at the concentration of that point. Once the calibration curve is created, then it is used to determine the concentration value of CSs. Solving for x_i, the equation becomes

$$x_i = (y_i \cdot f_i - \alpha)/(\beta \cdot g_i \cdot h_i) + e_i \qquad (4)$$

The random signal error term ε_i contributed by the signal during the calibration process is not retained. The term e_i is the random error on the concentration scale that typically is assumed to be normally distributed, but this distribution may vary in width over the concentration interval.

The concentration random error e_i contains several potential components, all of which are covered in the Clinical Laboratory Standards Institute document EP5-A3 *(3)*. Those components that are introduced within a calibration interval (on a specific instrument using specific reagents), such as repeatability, between-operator, between-run, and between-day imprecision, are equal to the signal random error ε_i from these sources divided by the slope of signal vs concentration on the calibration curve at position *i*. Thus, assuming a symmetric distribution:

$$[within\ calibration{:}]e_i = |\varepsilon_i/\beta| \qquad (5)$$

The calibration to calibration random error component of e_i is a function that modifies the underlying calibrator fit function f_i. The decision by the manufacturer to use a master calibration curve plus a small set of product calibrators (adjustors) vs using a larger set of product calibrators represents 2 ways to minimize this error component.

In the first case, a manufacturer may create the master calibration curve equation (curve shape) using multiple IVD medical devices and replicates over multiple runs for a specific reagent lot. The manufacturer will then value assign the 1- or 2-level calibrator set that adjusts this curve for each calibration in the laboratory. This error component comprises within-calibration variations on measurement of the small set of adjustors, medical device to medical device differences in curve shape from the master curve, changes in curve shape with reagent aging, and the inability of 1 or 2 adjusters to accurately correct for systematic shifts in the curve.

In the second case, a manufacturer uses multiple calibrator concentration levels to customize the calibration curve for any specific reagent lot of any age on any instrument. The primary source of variation is the within-calibration variation on measurement of each calibrator concentration. A recent report on the effect of using multiple calibrations to increase measurement trueness supports this point *(4)*. Using a mathematical model such as a linear regression or 4-parameter logistic curve can moderate the effects of this variation, but, as seen above, if this model is not a good fit to the actual kinetics of the MP, then f_i can drift away from systematic unity.

The remaining components of e_i are because of medical device to medical device and reagent lot-to-lot differences. These errors are not corrected by calibration and, therefore, are other sources of systematic error above and beyond error h_i caused by RM noncommutability. These sources of error must be kept in mind whenever generalized statements are made about commutability properties of an RM with specific MPs. However, the actual equation used to estimate CS values ignores all the systematic error terms and combines all the random error components into the single term

$$\hat{x}_i = (y_i - \alpha)/\beta + e_i \qquad (6)$$

When the systematic error terms are removed and disregarded by using this equation, different MPs provide different results for the same CS even if replication is used to reduce random error. Disregarding each systematic error function has different implications and effects. Different methods are used to reduce each of their magnitudes.

SYSTEMATIC ERRORS

The function *f* can influence MP results regardless of the type of calibration curve. If a linear fit is a good but not a perfect representation of the relationship between signal and concentration, the resultant calibration model fit and the resultant function *f* could look like the example in Fig. 1. In this case, a linear calibration fit is heavily weighted at low concentrations. If this function shape is seen in repeated calibrations, the manufacturer may wish to change the calibration model to, for example, a quadratic function rather than a linear function. If, however, the ratio function is not consistent from calibration to calibration, then the linear model may still be the best option. These considerations are valid whether multipoint calibrations are performed by the clinical laboratory when using an MP or whether a multipoint calibration is performed by the manufacturer and 1- or 2-point adjustments are made by the clinical laboratory.

The function *g* is influenced by the process of preparing the MP calibrators. A common approach begins with a high concentration standard preparation of analyte that can be diluted gravimetrically or volumetrically with matrix containing zero analyte. The resulting serially diluted series of calibrator samples containing proportionally related fractional amounts of analyte compared with the initial concentrated material typically produces a highly linear relationship between amount of measurand and fractional dilution. In other cases, if, for example, an equilibrium must be reached between free and bound analyte, a distinctly nonlinear relationship

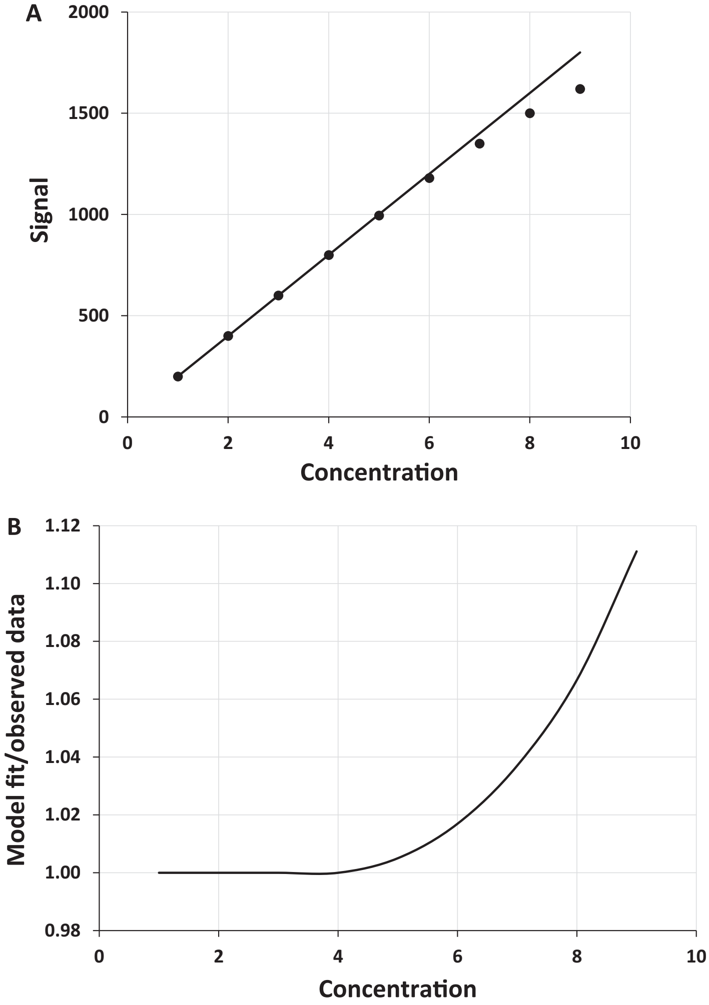

Fig. 1. Example systematic error over the measuring interval owing to a calibration scheme that does not capture the true relationship between signal and measurand value.

(A) shows the calibration fit of the observed data. The line is the calibration model fit, and the points are the observed calibration data. (B) shows the function *f* for the observed calibration fit.

may be observed, and other means must be used to determine the amount of measurand in each calibrator sample in the dilution series. A calibrator dilution series that provides proportional results (i.e., $f = 1$), by definition, also provides linear results (intended vs actual levels can be fit to a straight line). If an MP using proportionally related calibrators does not provide the correct concentration of a commutable RM or does not mimic the results for CSs from a reference MP, then the concentration of the initial high concentration standard preparation of the analyte, and, thus, the dilutions to prepare product calibrators, may be adjusted by a multiplier to provide true values over the entire measuring interval (i.e., setting $g = 1$).

Just because a calibrator dilution series is linear does not mean its results are proportional. This discrepancy can occur if the serial dilutions are not prepared correctly or if the dilution matrix contains some small amount of the analyte. In these cases, there may be need for an offset at zero concentration that translates into an increased actual to intended amount of substance ratio at low concentrations. This discrepancy is shown in the example in Fig. 2. In this example, the bias from the true concentration at the highest concentration is about −13%, but the contribution of the dilution matrix moves this actual to intended concentration ratio higher at lower concentrations.

Even if calibration provides proportionality ($f = 1$) and the calibration system has been aligned to match

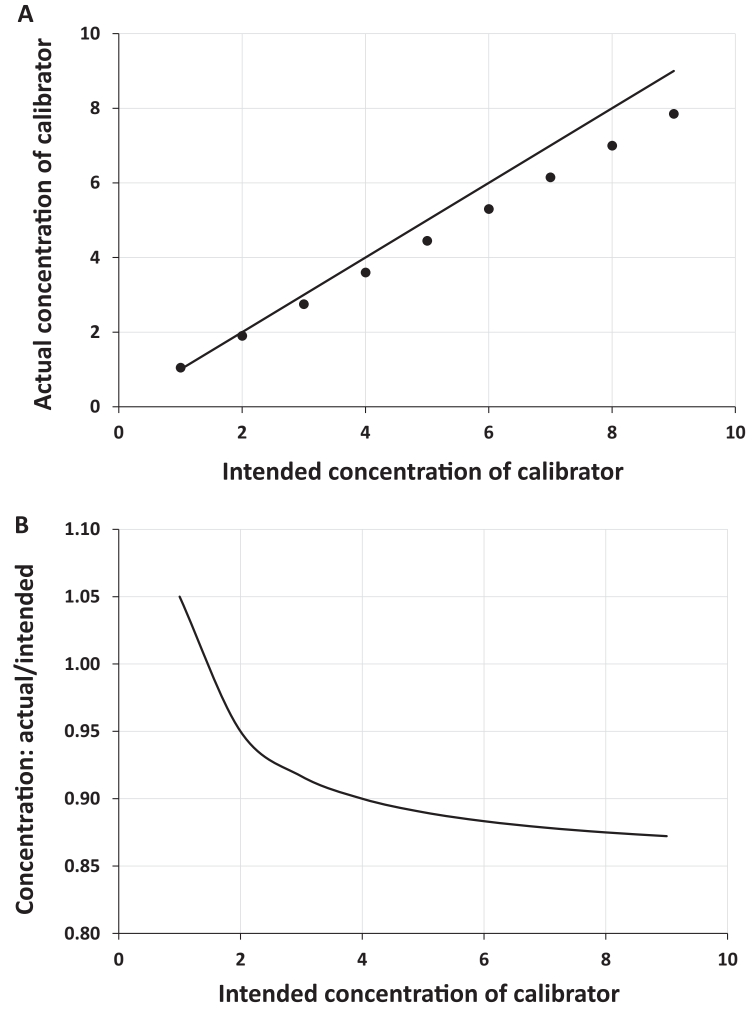

Fig. 2. Example systematic error from a serial calibrator dilution process that provides actual amounts of measurand that are not proportional to the intended amounts.

(A) shows the calibration fit of the calibrator dilution series. The line is unity for a one-to-one relationship between actual and intended concentrations of the calibrators, and the points are the observed concentrations of the diluted calibrator series. (B) shows the function *g* for the actual and intended concentrations.

the value of the RM ($g = 1$), the RM may not be commutable with an MP ($h \neq 1$) with the effect that dilutions of the RM made to match the concentrations of CSs do not give the same MP response as for the CSs containing the same amount of measurand. This noncommutability phenomenon is difficult to measure directly. A common way to visualize it is to compare results for CSs and RMs intended for use as calibrators between ≥2 different MPs using a plot of results for CSs and RMs *(5)*. As with the other 2 functions (*f* and *g*) mentioned above, any bias owing to noncommutability (function *h*) may be constant across the measuring interval or may vary in magnitude with concentration.

Commutability Assessment by the Calibration Effectiveness of a Reference Material

GENERAL CONSIDERATIONS

Manufacturers of an RM must ensure that their material is fit for the purpose of being a higher order calibrator in the traceability chain of lower order product calibrators from multiple MPs. MPs can be excluded from the comparison if

- MP results have high imprecision (*e*),
- A poor fitting mathematical model or a suboptimal fitting algorithm is used for calibration (function $f \neq 1$), or
- The MP is relatively sensitive to individual sample-specific interferences.

For the commutability assessment, the baseline assumptions are that when a single RM is used in the calibration hierarchy chain to recalibrate each medical laboratory MP, that

- The RM will be commutable compared with CSs (function $h = 1$), and
- The resultant rescaling will align the clinical laboratory MP product calibrators to their correct values (function $g = 1$).

If these assumptions are met, then recalibrated clinical laboratory MPs will give equivalent CS results, within an acceptability criterion, across the concentration interval and, subsequently, the RM can be said to be commutable. If some MPs give different results after recalibration with the RM, then remaining sources of error must be investigated before concluding that the RM is noncommutable for those MPs.

The following example describes a method of assessing the commutability of an RM (function *h*) across multiple MPs using a set of CSs. A requirement for the MP to MP variability for CS results is specified (e.g., bias range $\leq$ 6%) as the criterion for commutability. This variability is measured as the intermeasurement procedure bias range (IMPBR). The accurate determination of this parameter depends on the robust determination of bias for each MP regardless of imprecision or interferences. The underlying assumption is that a single bias estimate can be provided for each MP (i.e., bias does not depend on concentration). This assumption should be tested. In the following description, an example data set is used to illustrate the process of assessing RM commutability by the RM's effectiveness to improve the agreement of results for a set of CSs after recalibration of the MPs with the RM. The analyses described assume $\geq$32 CSs *(6)* and $\leq$20 MPs *(7)*. The cited articles can be referenced for different order statistic options if fewer CSs or more MPs are used.

MEASUREMENT PROCEDURE SCREENING

A set of CSs is obtained per specifications [see part 1 of this series *(2)*] that reasonably covers the expected concentration interval of the candidate RMs to be assessed for commutability. Each CS and RM level provided is measured using each MP with sufficient replication to meet the uncertainty requirements for the study. The replication for each of the small number of RM levels will be more extensive than for each of the many CSs with the RMs distributed in several positions among the CSs in the sequence of measurements by each MP. In addition, the testing sequence for CSs should be random regarding concentration. An example of a detailed experimental design is given in part 2 of this series *(8)*. The concentration error term e_i and its variance components must be considered for each MP in designing and performing such a study. The mean of the replicate results for each CS is used in subsequent calculations for comparing the agreement among results for various MPs.

Table 1 in the online Data Supplement shows results from 40 CSs for 7 MPs (labeled MP1–MP7) using each MP's manufacturer-specified calibration scheme plus an across-MP target result. Given an assumption of normality across MP results for each CS and the number (n) of MPs is <20, the most robust estimate of central tendency (target result) is the trimmed mean (T) *(7)* of the MP results for each CS (X = ordered results) computed as

$$T = \frac{\sum_{i=2}^{n-1} X_{(i)}}{n-2} \qquad (7)$$

Other options for estimating T are available for situations when data are censored or the underlying distribution cannot be assumed to be normal *(9)*. Alternative methods have been described in efforts to harmonize results across MPs *(10–12)*. Note that the sample identifiers (IDs) in Table 1 of the online Data Supplement are ranked in order of increasing trimmed mean (target result) concentration.

These data can also be represented by a plot with the percent difference from the trimmed mean target plotted against the sample ID, as seen in Fig. 3. The sample ID is related to sample concentration value, with higher sample IDs having higher concentration values (see Table 1 in the online Data Supplement for concentrations). This plot is the best way to see whether specific samples behave differently (e.g., because of interferences) across different MPs because the IDs are aligned and evenly spaced. Scaling in this way directly displays overall bias for each MP at each CS. Viewing this plot, it is seen that MP1 results are close to the trimmed mean target, whereas the other MS results (labeled MP2–MP7) range from approximately 50% higher to −30% lower than the target results. The actual median bias differences are listed in Table 1. If desired, median CIs may be determined per Clinical Laboratory Standards Institute document EP09 *(13)*.

Sample 14 gives outlying results for most of the MPs, which could be used as a reason to exclude this sample from this analysis or from future testing. Sample 7 gives a high outlying result for MP3, indicating a potential interference specific to that sample on MP3. These 2 samples are retained in the following analysis because the order statistics used are robust enough to not be overly

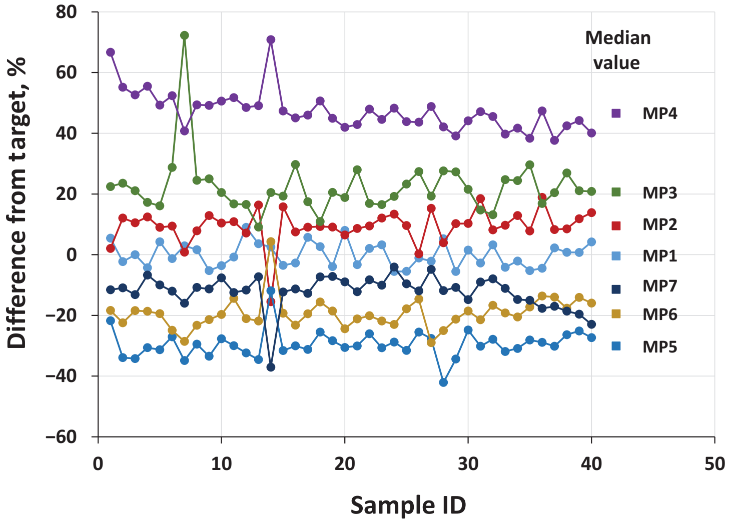

Fig. 3. Difference in percent from the target value (trimmed mean) for each of 40 CS original results from 7 MPs. The median difference for each MP is shown as the square symbol, along with the MP identifiers to the right.

influenced by such outlying results. Because of the outlying result, MP3 appears to have higher imprecision than the other MPs, which is seemingly confirmed by measuring the SD of the percent differences over all CSs (9.5% for MP3, 4.0% for MP1, and 5.7% for MP2). However, a more robust variability estimate called the quasi-range *(6)* $W_{(3)}$ scaled to MP1 gives 5.9% for MP3, 4.0% for MP1, and 5.6% for MP2 (see Table 2 here and the online Data Supplement for a description of this technique). Therefore, there is no compelling reason to eliminate MP3 from consideration in this analysis. Such measures of imprecision could be used, however, to determine whether more replication is warranted for selected MPs.

MP7 results have a consistent offset from the target except at the highest concentrations for which the bias changes with concentration. MP4 has a steadily increasing bias as results get lower in concentration. Except for the trends noted for MP4 and MP7, all other MPs have a relatively consistent percent difference from target values. When trends are apparent, as with MP4 and MP7, the median may not provide accurate estimates of overall bias. This situation will be addressed later during considerations for recalibration.

Table 1 (Measurement procedure screening) presents additional calculations of the median percent bias (from target) over all 40 samples for each MP for the screening results. In addition, the IMPBR over the 7 MPs is also provided. This IMPBR is the value that can be compared with the maximum cross-MP commutability criterion described above as 6%. Because the MPs are not yet standardized to the RM, the resultant high IMPBR of 76.6% is not unexpected, indicating that the next steps in the process should be followed.

USING A REFERENCE MATERIAL FOR CALIBRATION TRACEABILITY

The next step in the commutability assessment is to determine whether the RM(s) when used for recalibration

Table 1. Median percent biases across all CSs for each MP and IMPBR across all MPs.

	MP1	MP2	MP3	MP4	MP5	MP6	MP7	Bias range (IMPBR)
Results from measurement procedure screening								
Median bias, %	−0.4	9.5	20.7	46.6	−30.0	−19.4	−11.4	76.6
Results from commutability assessment after recalibration								
Median bias, %	−0.8	−0.3	−0.2	−0.3	0.4	−21.3	−1.1	21.6
Median bias excluding MP6, %	−0.8	−0.3	−0.2	−0.3	0.4	Exclude	−1.1	1.5

Table 2. Measures of sample to sample variation of percent differences from the target value for each MP.

MP	MP1	MP2	MP3	MP4	MP5	MP6	MP7
SD before calibration, %	4.01	5.72	9.54	6.73	4.59	5.30	5.65
Adjusted $W_{(3)}$ before calibration, %	4.01	5.57	5.90	5.89	3.49	3.96	4.63
SD after calibration, %	4.05	4.99	7.83	3.80	6.34	4.89	6.45
Adjusted $W_{(3)}$ after calibration, %	4.05	4.98	4.95	3.45	4.39	3.57	5.06

SD for each MP over all CSs.

of the MPs can reduce the IMPBR. The recalibration effort must consider both the random errors inherent in the calibration process ε_i and the potential lack of fit described by the function f_i. The plot from Fig. 3 can be reviewed to determine whether a single concentration level of RM can be used to recalibrate the MPs. The imprecision seen within the CS and RM samples could be used to determine whether enough replication was used to meet the uncertainty requirements of the study on each MP. As noted, most of the MPs have a consistent proportional offset from the target values. This offset was measured with a least-squares linear regression of MP bias vs trimmed mean for MP1, MP2, MP3, MP5, and MP6 with all P values of the slope $\geq$0.17. This P value implies that for those MPs a single RM level can be used to proportionally correct the bias from the target (i.e., set $g = 1$).

MP4 showed an increasing bias at lower concentrations (slope P value < 0.0001). Such a change over concentration can be caused by a poor calibration fit ($f \neq 1$) or errors in creating calibration material ($g \neq 1$). The shape of the MP4 bias plot implies that there is a constant offset in addition to a proportional offset. These offsets could be corrected by using a set of RMs with at least 2 concentrations of the measurand. Given the relatively linear slope of MP4 bias over the concentration interval, such a strategy could bring it into alignment with the rest of the MPs. However, it would not address the potentially poor design and implementation process used in creating this MP. It is recommended that the RM provider describe to the MP manufacturer how their MP results differ from the other MP results so that the manufacturer can investigate potential design issues.

Further investigation may be required to determine the path forward for MP7 (slope P value = 0.0497). Typically, the process of making calibration material affects low concentration samples more than high concentration samples, as seen with MP4, so function g is unlikely to play a role here. This observation is more likely to be an issue of poor calibration fit ($f \neq 1$), which is a function of MP design (see Fig. 1). Therefore, although such variation could potentially be corrected by providing 3 concentration levels of the RM, the RM manufacturer should not be expected to solve individual MP design issues. Again, the RM provider should describe to the MP manufacturer how their MP results differ from the other MP results so that the manufacturer can investigate potential design issues. MP3 had a large outlying result for sample 7, suggesting a sample interferent different from any other MP. The RM provider can decide whether to drop sample 7 from consideration with MP3 or to drop MP3 from the list of MPs to be assessed for commutability. Regardless of this decision, the MP manufacturer should be notified of the identified issue.

EFFECTIVENESS OF RECALIBRATION WITH THE RM AS AN ASSESSMENT OF COMMUTABILITY

The general approach is to substitute the candidate RM to be assessed for commutability for the RM currently used by an MP manufacturer in the MP's calibration traceability scheme. In other words, the RM to be assessed for commutability should be used directly as a calibrator for the clinical laboratory MP or as a calibrator for a manufacturer's value transfer procedure in its traceability scheme as described in ISO 17511 *(1)*. The set of CSs are then remeasured with each MP now having its calibration traceable to the candidate RM.

Table 2 in the online Data Supplement shows results from the same 40 CSs after the 7 MPs (labeled MP1C–MP7C) have been recalibrated using the candidate RM. The option of using a 2-level RM was chosen for all MPs to correct for the bias seen in MP4 and any residual constant bias not identified by the difference plots for the other MPs. These adjusted product calibrator values could be used to recompute the concentrations from the signals obtained from the original CS measurements without repeating the experiment. Alternatively, this adjustment could be made to the product calibrator levels and the CS then measured again using the MP with the new calibration. The first approach is preferred to avoid incremental measurement error introduced by the second approach.

The results for the set of CSs after recalibration can be represented by a difference plot as seen in Fig. 4. Before discussing the plots in Fig. 4, it is useful to review the assumptions that underlie the analyses that have been

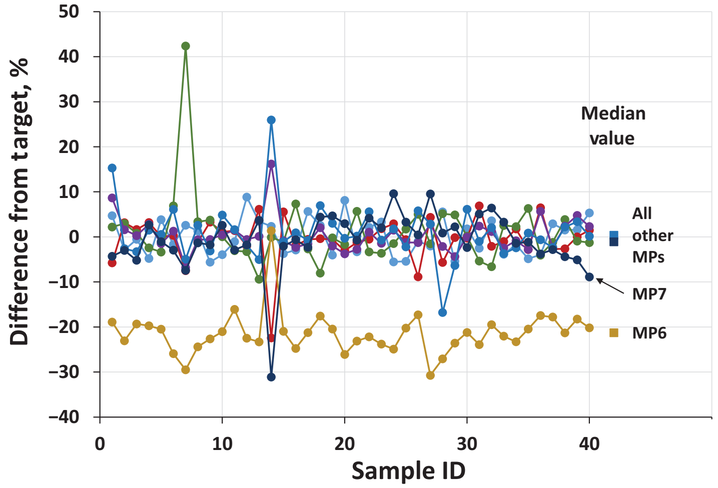

Fig. 4. Difference in percent from the target value for the same 40 CS results after recalibration of the 7 MPs shown in Fig. 3 with traceability to the RM.

The median difference for each MP is shown as the square symbol, along with the MP labels to the right. Note that the MP6 median is separated from the other medians that are difficult to distinguish from each other. The MP colors are the same as in Fig. 3. MP7 is the dark blue symbol as pointed out on the figure.

done. First, the use of robust measures of bias has reduced the effects of random variation and sample-specific interferences on the measurements of bias. Second, systematic error can be completely described for each MP by the functions *f* (calibration fit), *g* (calibrator value assignment), and *h* (RM commutability). Third, by definition, the use of the RM for calibration traceability sets function $g = 1$. Fourth, the use of a 2-level RM will correct for errors in functions *f* even if they are offsets that change in a linear fashion over concentration. Fifth, any remaining bias is owing to noncommutability (function *h*) or the inability of a linear correction to adjust for calibration lack of fit (function *f*).

The plots in Fig. 4 indicate that for MP1 through MP5 and MP7 the median biases are close to zero. Given the above assumptions and lack of slope for MP4, it can be concluded that for MP1 through MP5 the recalibration using the RM set the function $g = 1$ and adjusted for any issues with calibration fit (function *f*). In addition, the RM was commutable with CS for MP1 through MP5 (function $h = 1$). The MP6 median bias is different from zero. For MP6, the bias is consistent over the concentration interval. Therefore, because $g = 1$ and there are no indications of inconsistent bias over the concentration interval, then $f = 1$ as well. Assuming sources of random error have been reduced by study design, the remaining source of error is the systematic error function *h*, meaning that the RM is not commutable with MP6.

MP7 is problematic. Although its median percent difference from target is close to zero, the linear adjustment did not eliminate the decreasing bias trend at the highest concentrations. Therefore, the bias is not consistent across the concentration interval and the median is not an accurate description of overall bias for MP7. The inconsistent bias is presumably because $f \neq 1$ at higher concentration. It is an open question whether this bias is because of a consistent systematic calibration error or the specific calibration(s) used in this study. The effect of the calibration to calibration random effect could be reduced by increasing the number of calibrations *(4)*.

The bias analysis was performed on the data in Table 2 in the online Data Supplement with the results shown in the second part of Table 1 here (Commutability assessment after recalibration). Again, the IMPBR is compared with the predetermined criterion. In this case, the estimate of 21.6% still does not meet the commutability requirement of IMPBR ≤ 6%. However, the median MP to MP variability has been notably reduced (i.e., from 76.6%). The median bias for MP6 is much larger than the other MPs. This excess bias indicates that the RM is not suitable for use with MP6. This observation does not necessarily mean the RM has a problem, but simply that it is not commutable for MP6 and cannot be used to provide calibration traceability for MP6. Therefore, MP6 should be excluded from the assessment of RM suitability. After excluding MP6, the bias range reduces to 1.5%, which meets the 6% IMPBR criteria.

Recalibration will not change the overall sample to sample variability. Evidence for this statement is shown in Table 2, which, for each MP, presents the SD of the percent differences over all CSs and the quasi-range $W_{(3)}$ adjusted to the MP1 SD (see the online Data Supplement for a discussion of this technique). The only exception is MP4, for which before recalibration $W_{(3)}$ was 5.89% and afterward was 3.45%. This change is because before recalibration the variability estimate was made over a bias that changed with concentration. After recalibration, this change in bias over concentration was eliminated and the variability was subsequently reduced.

Recalibration with RM also cannot solve the incidence of measurement interferences seen in some individual CSs with some MP results, nor can the 2-level RM chosen in this example solve the nonlinear issues seen for MP7. An RM with a different concentration level or additional RMs may be considered to address the nonlinearity ($f \neq 1$) seen for MP7. Alternatively, it is necessary for the manufacturer of MP7 to improve the measurement performance. It is not the responsibility of the RM manufacturer to solve the problems of IVD manufacturers. However, it is important for RM manufacturers to work collaboratively with IVD manufacturers to ensure the RM(s) will be generally suitable for use. The MP6 manufacturer should be notified that their results were excluded to reach the IMPBR specification and the RM was not commutable for use with MP6.

Conclusion

A process of assessing the commutability of an RM has been described that uses a set of CSs to determine bias between MPs. Robust methods of determining bias (i.e., difference from trimmed mean of all MP results) are used to reduce the effects of imprecision and sample-specific interferences. The sources of bias have been elucidated by describing their various effects on the calibration process used by each MP to provide CS results. The sources of bias include lack of calibration fit, calibrator bias, and noncommutability. When a candidate RM is used as the highest-level material in the traceability scheme for each MP, the resultant CS bias estimates can help separate the overall bias into these constituent sources of error. This process can identify those MPs whose bias can be eliminated through use of the candidate RM, those MPs whose design may need to be updated, and those MPs for which the RM is noncommutable with CSs.

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. ISO 17511:2003. In vitro diagnostic medical devices–measurement of quantities in biological samples–metrological traceability of values assigned to calibrators and control materials. Geneva (Switzerland): ISO; 2003.
2. 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.
3. Evaluation of precision of quantitative measurement procedures; approved guideline. 3rd Ed. CLSI document EP05-A3. Wayne (PA): Clinical and Laboratory Standards Institute; 2014.
4. Akbas N, Budd JR, Klee GG. Multiple calibrator measurements improve accuracy and stability estimates of automated assays. Scand J Clin Lab Invest 2016;76:177–80.
5. Evaluation of commutability of processed samples; approved guideline. 3rd Ed. CLSI document EP14-A3. Wayne (PA): Clinical and Laboratory Standards Institute; 2014.
6. Cadwell JH. The distribution of quasi-ranges in samples from a normal population. Ann Math Statist 1953;24:603–13.
7. Dixon WJ. Estimates of the mean and standard deviation of a normal population. Ann Math Statist 1957;28:806–9.
8. Nilsson G, Budd JR, Greenberg N, Delatour V, Rej R, Panteghini M, et al. IFCC working group recommendations for assessing commutability part 2: using the difference in bias between a reference material and clinical samples. Clin Chem 2018;64:455–64.
9. David HA, Nagaraja HN. Order statistics. 3rd Ed. New York (NY): Wiley; 2003.
10. Van Houcke SK, Van Aelst S, Van Uytfanghe K, Thienpont LM. Harmonization of immunoassays to the all-procedure trimmed mean–proof of concept by use of data from the insulin standardization project. Clin Chem Lab Med 2013;51:e103–5.
11. Stockl D, Van Uytfanghe K, Van Aelst S, Thienpont LM. A statistical basis for harmonization of thyroid stimulating hormone immunoassays using a robust factor analysis model. Clin Chem Lab Med 2014;52:965–72.
12. Thienpont LM, Van Uytfanghe K, De Grande LAC, Reynders D, Das B, Faix JD, et al. Harmonization of serum thyroid-stimulating hormone measurements paves the way for the adoption of a more uniform reference interval. Clin Chem 2017;63:1248–60.
13. Measurement procedure comparison and bias estimation using patient samples; approved guideline. 3rd Ed. CLSI document EP09-A3. Wayne (PA): Clinical and Laboratory Standards Institute; 2013.